




3 1761 10374371 2



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743712>



SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2002

•

VOLUME 28

•

NUMBER 1



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2002 • VOLUME 28 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2002

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

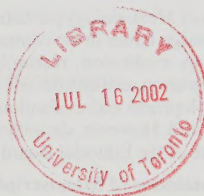
June 2002

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Past Chairman)

E. Rancourt (Production Manager)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidirolou, *Statistics Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*
G. Nathan, *Hebrew University, Israel*

D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

SURVEY METHODOLOGY
A Journal Published by Statistics Canada
Volume 28, Number 1, June 2002

CONTENTS

In This Issue	1
Waksberg Invited Paper Series	
WAYNE A. FULLER	
Regression Estimation for Survey Samples	5
Special Section "Remembering Leslie Kish"	
GRAHAM KALTON	
Leslie Kish's Impact on Survey Statistics	25
LESLIE KISH	
New Paradigms (Models) for Probability Sampling	31
CHARLES H. ALEXANDER	
Still Rolling: Leslie Kish's "Rolling Samples" and The American Community Survey	35
JEAN-MICHEL DURR and JEAN DUMAIS	
Redesign of the French Census of Population	43
Regular Papers	
IAN CAHILL and EDWARD J. CHEN	
Benchmarking Parameter Estimates in Logit Models of Binary Choice and Semiparametric Survival Models	51
STEVEN T. GARREN and TED C. CHANG	
Improved Ratio Estimation in Telephone Surveys Adjusting for Noncoverage	63
YVES TILLÉ	
Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement	77
JUN SHAO and SHAIL BUTANI	
Variance Estimation for the Current Employment Survey	87
MICHAEL P. COHEN	
Implementing Rao-Shao Type Variance Estimation with Replicate Weights	97
RICHARD VALLIANT	
Variance Estimation for the General Regression Estimator	103

In This Issue

This issue of *Survey Methodology* contains the second in an annual invited paper series in honour of Joseph Waksberg. A brief description of the series and a short biography of Joseph Waksberg were given in the June 2001 issue of the journal. The author of the Waksberg Invited Paper for 2002 is Wayne Fuller. I would like to thank the members of the Committee, Graham Kalton (chair), Chris Skinner, David Binder and Paul Biemer, for having chosen such a distinguished statistician, who has made profound contributions to many areas of statistical theory and practice, as the author of the second paper in the Waksberg Invited Paper Series.

In his paper entitled "Regression Estimation for Survey Samples" Wayne Fuller presents a broad overview of historical and recent developments in the use of regression models in surveys for estimation, weight calibration and non-response adjustment. After a brief introduction and historical background, he discusses the use of regression models for estimation in complex surveys from a design based perspective. He follows this with an exploration of the model based perspective. Other topics discussed are the use of regression models for multinomial data, techniques available when auxiliary variables are available for every unit of the population, and regression to account for the effects of non-response in surveys. Finally, consideration of a few practical aspects of applications rounds out this insightful overview of an important area of inference from survey data to which Wayne Fuller himself has made many important contributions.

This issue also contains a special section "Remembering Leslie Kish" which includes four papers, one by Leslie Kish himself containing some of his last thoughts on the topics of combining samples and surveys. Two of the other papers discuss implementations of Leslie Kish's idea of rolling censuses. These two papers were also presented at the Statistics Canada Symposium 2001 in a special session entitled "Remembering Leslie Kish".

The first paper in the special section, by Graham Kalton, presents an inspiring overview of Kish's contributions to many areas of statistics. Many of the problems that Kish worked on are put into historical perspective and their practical importance is emphasized.

The paper by Kish presents ideas that he was still working on at the time of his death in October 2000. I am grateful to Graham Kalton and Jack Gambino for making editorial corrections to the paper, but it is presented largely as it was at the time of Kish's death. In this paper he argues that, just as statistics represented a new paradigm in the scientific method, and survey sampling required a new paradigm in statistics, so rolling samples and multi-population surveys require new paradigms in survey methods. We can only speculate as to what the final paper would have been like had Kish lived.

Alexander describes the American Community Survey, planned to be introduced by the U.S. Census Bureau in coming years as a replacement for the decennial census long form. This is a very large survey based very much on the idea of rolling samples and censuses that Kish introduced more than twenty years ago. This paper discusses the concepts, frame, sampling design, and cumulation of samples and weighting.

The final paper in the special section, by Durr and Dumais, describes the new rolling census being introduced in France to replace their more traditional census. In this rolling census, every small commune will be surveyed once within a five year period; larger communes will be divided into five rotation groups, each rotation group being surveyed in one of the five years. This paper describes objectives, design and estimation procedures for the rolling census.

In their article, Cahill and Chen develop an approach to exploit data from multiple surveys and epochs by benchmarking the parameter estimates of logit models of binary choice and semi-parametric survival models. Estimates obtained from a survey rich in explanatory variables are benchmarked to information from a survey with significant historical depth. Cahill and Chen demonstrate how the method can be applied, using the maternity leave module of the LifePaths dynamic microsimulation project at Statistics Canada.

Garren and Chang consider the problem of the non-telephone population in telephone surveys using random digit dialing. Using Public Use Microdata Samples, the propensity that a household owns a phone is estimated using generalized linear regression and is used during estimation. Asymptotic biases and variances are presented for both the non-poststratified and poststratified estimators incorporating and not incorporating the estimated propensity. These four estimators are further compared through a simulation study.

The article by Tillé develops an estimator that can be used to avoid the problem of empty post-strata that can occur with the usual post-stratified estimator. The idea involves using a conditionally weighted estimator and conditioning on ranks in the population of an auxiliary variable known for all units of this population. In this way, the sizes of the post-strata are set in the sample and random in the population. The next step is to calculate the mean of the conditionally weighted estimators to obtain greater stability. The estimator obtained is calibrated on distribution, linear and exactly unbiased. A simulation study is used to show that the proposed estimator is more robust than the generalized regression estimator when the relation of the variable of interest and the auxiliary variable is not linear. Lastly, the article proposes an approximate estimator of the variance verified using simulations.

Shao and Butani consider the problem of estimating variances for imputed survey estimators. They show that the resulting variances can be estimated in two parts, the first of which can be estimated using a grouped half-sample method that incorporates adjustments to take imputation into account. As the estimation of the second part may entail many derivations, Shao and Butani propose an adjustment to the grouped half-sample method that leads to approximately unbiased variance estimates.

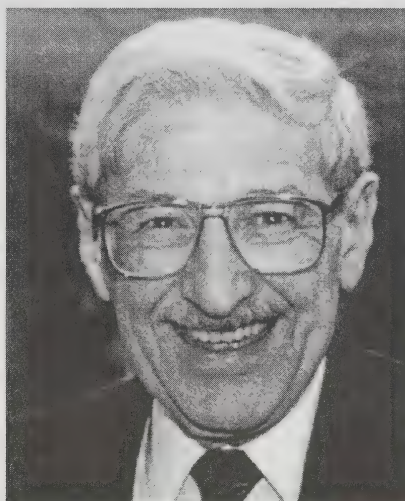
In his paper Cohen describes a method to implement Rao and Shao's jackknife method of estimating variances to account for imputation using replicate weights. Rao and Shao's method involves calculation, for each jackknife replicate, adjusted values of imputed data points. The method can be used with either mean imputation or hot deck imputation. Cohen's method involves adding extra rows to the replicate weight file. For each imputed value, one extra row is added for each respondent in the same imputation class.

In the last paper of this issue, Valliant studies several variance estimators for the General Regression (GREG) estimator. The interest is in finding variance estimators that, under certain conditions, are approximately unbiased for both the design-variance and the model-variance even if the model that motivates the GREG has an incorrect variance parameter. A key feature of these robust estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. It is shown that the delete-one jackknife implicitly includes the leverage adjustments and is a good choice from either the design-based or model-based perspective. A simulation study shows that these variance estimators have small bias and produce confidence intervals with near-nominal coverage rates.

M.P. Singh

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association*. The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*.



JOSEPH WASKBERG

2002 WAKSBERG INVITED PAPER

Author : Wayne A. Fuller

Wayne A. Fuller is Emeritus Distinguished Professor in Statistics and Economics at Iowa State University. He has published approximately 100 articles in more than twenty journals and is author of the texts *Introduction to Statistical Time Series* and *Measurement Error Models*. As a member of the Survey Group at Iowa State University, he had primary responsibility for developing estimation procedures for a large longitudinal national survey called the *U.S. National Resources Inventory*. His research interests in survey sampling include regression estimation, small area estimation, imputation, and multiple phase sampling. He currently chairs the Advisory Committee on Statistical Methods of Statistics Canada.

MEMBERS OF THE WASKBERG PAPER SELECTION COMMITTEE (2002-2003)

David A. Binder (Chair), *Statistics Canada*
J. Michael Brick, *Westat, Inc.*
David R. Bellhouse, *University of Western, Ontario*
Paul Biemer, *Research Triangle Institut, U.S.A.*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)

Past Authors:

Gad Nathan (2001)

Nominations:

Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, D.A. Binder, at Statistics Canada, 3rd, floor R.H. Coats Bldg. Tunneys' Pasture, Ottawa, Ontario, Canada, K1A 0T6, by e-mail binderdav@statcan.ca or by fax (613) 951-5711. Nominations and suggestions for topics must be received by December 6, 2002.

Regression Estimation for Survey Samples

WAYNE A. FULLER¹

ABSTRACT

Regression and regression related procedures have become common in survey estimation. We review the basic properties of regression estimators, discuss implementation of regression estimation, and investigate variance estimation for regression estimators. The role of models in constructing regression estimators and the use of regression in nonresponse adjustment are explored.

KEY WORDS: Auxiliary information; Calibration; Least squares; Design consistency; Linear prediction.

1. INTRODUCTION

Design and estimation in survey sampling involve the use of information about the study population to construct efficient procedures. While design and estimation are intimately related, with estimators depending on the design, the two topics are often treated somewhat separately in the survey sampling literature. We follow tradition first studying estimation treating the design as given. The estimation task is to combine the available information about the population, with the sample data to produce good representations of characteristics of interest.

Regression estimation is one of the important procedures that use population information or information from a larger sample, to construct estimators with good efficiency. The information, sometimes called *auxiliary information*, may have been used in the design or may not have been available at the design stage. In surveys of the human population, the information often comes from official sources such as the national census. Similar sources may provide information for other types of surveys. For example, in a survey of land use the total surface area, the area owned by the national government, and the area in permanent water bodies may be available from national data archives.

Three distinct situations can be identified with respect to the nature of the auxiliary information that is available. In the first, the values of the auxiliary vector \mathbf{x} are known for each element in the population at the time of sample selection. In this case the auxiliary variable can be used in designing the sample selection procedure.

In the second situation all values of the vector \mathbf{x} are known, but a particular value cannot be associated with a particular element until the sample is observed. In this case, the auxiliary information cannot be used in design, but a wide range of estimation options are available once the observations are available. For example, the population census may give the age-sex distribution of the population, but a list of individuals and their characteristics is not

available to non governmental institutions selecting samples.

In the third situation, only the population mean of \mathbf{x} is known, or known for a large sample. In this case, the auxiliary information cannot be used in design and the estimation options are limited. For example the U.S. Department of Agriculture might release an estimate of the total number of animals of a particular type on farms on a particular date. Our discussion concentrates on this situation.

Two estimation situations can also be identified. In one, a single variable and a parameter, or a very small number of parameters, is under consideration. The analyst is willing to invest a great deal of effort in the analysis, has a well formulated population model, and is prepared to support the estimation procedure on the basis of the reasonableness of the model. In the second situation, a large number of analyses of a large number of variables is anticipated. No single model is judged adequate for all variables. The prototypical example of the second situation is the case in which a data set is prepared by the survey sampler to be analyzed by others. Because the person preparing the data set does not have knowledge of the analysis variables, emphasis is placed on the use of estimators that can be defended with minimal recourse to models.

Regression estimators fall in the class of linear estimators. Linear estimators have a particular advantage in survey sampling because once the weights are calculated they are appropriate for any analysis variable. Several properties of estimators will be examined in our discussion. Given a model, we accept the classical goal of minimizing the mean square error in a class of estimators. That class may be the class of linear estimators that are unbiased under the model, but the class may be further restricted.

Estimators that are scale and location invariant can be used in general settings. Mickey (1959) suggested that the term regression estimator be restricted to linear estimators that are location and scale invariant. While we may not adhere strictly to this definition, we support the distinction

¹ Wayne Fuller, Emeritus Distinguished Professor, Iowa State University, 221 Snedecor Hall, Ames, IA 50011-1210, U.S.A.

between estimators that are location and scale invariant and those that are not. We consider location invariance to be important for sampling designs where the unit of interest for analysis is also the sampling unit. For cluster and two stage designs in which weights are constructed for primary sampling unit totals, location invariance is less important.

Models play an important role in the construction of regression estimators. It is desirable that the estimators retain good properties if the model specification is not exact. Therefore properties conditional on the realized finite population, as well as properties under the model, are important.

Linear estimators that reproduce the known means of the auxiliary variables are said to be calibrated. This is a desirable property in that, for example, the marginals of tables with an auxiliary variable as an analysis variable agree with known totals. If the auxiliary variable is of no analytic interest, then calibration is less important.

2. BACKGROUND

The earliest references to the use of regression in survey sampling include Jessen (1942) and Cochran (1942). Regression in similar contexts would certainly have been used earlier and Cochran (1977, page 189) mentions a regression on leaf area by Watson (1937). It is interesting that Jessen's use of regression was essentially composite estimation where regression was used to improve estimates for two time points given samples at each point with some common elements in the two samples. Cochran (1942) gave the basic theory for regression in survey sampling relying heavily on linear model theory. He showed that the linear model did not need to hold in order for the regression estimator to perform well. He derived an expression for the $O(n^{-1})$ bias and an $O(n^{-2})$ approximation for the variance. He also showed that for the model with regression passing through the origin and error variances proportional to x , the ratio estimator is the generalized least squares estimator.

Regression estimation attracted theoretical interest in the 1950's, often in the form of studies of the bias. See Mickey (1959). Brewer (1963) is an early reference that considers linear estimation using a superpopulation model to determine an optimal procedure. He was concerned with finding the optimal design for the ratio estimator and discussed the possible conflict between an optimal design under the model and a design that is less model dependent. See also Brewer (1979). Royall (1970) argued for the use of models, that the conditional properties that are important are those conditional on the auxiliary information in the sample, and that the design should be chosen to optimize those properties. Royall and his coworkers, *e.g.*, Royall and Cumberland (1981), studied the conditional properties of regression estimators, conditional on the realized sample of auxiliary variables.

A great deal of research was conducted in the 1970's and 1980's on the general nature of the regression estimator in survey samples and on the degree to which the model prediction approach can be reconciled with the design perspective. Fuller (1973, 1975) gave the large sample properties of a vector of regression coefficients computed from a survey sample. Isaki (1970) studied regression estimators and the results were published in expanded versions in Isaki and Fuller (1982) and Fuller and Isaki (1981). It was shown that a regression estimator constructed under a model is design consistent for the population mean if the model contains certain variables. Cassel, Särndal and Wretman (1976) considered both model and design principles in estimator construction and suggested the term "generalized regression estimator" for design consistent estimators of the total of the form

$$\hat{T}_{y,\text{GREG}} = \hat{T}_{y,\text{HT}} + (T_{x,N} - \hat{T}_{x,\text{HT}})\hat{\beta},$$

where $\hat{T}_{y,\text{HT}}$ and $\hat{T}_{x,\text{HT}}$ are the Horvitz-Thompson estimators of the totals of y and x , respectively, $T_{x,N}$ is the known population total of x and $\hat{\beta}$ is an estimated regression coefficient. Särndal (1980), Wright (1983), and Särndal and Wright (1984) discussed classes of regression estimators. The text by Särndal, Swensson and Wretman (1992) contains an extensive discussion of regression estimation and Mukhopadhyay (1993) is a review.

It was the 1970's before the use of regression for general purpose, multiple characteristic, surveys appeared and it was the 1990's before the use of regression weighting could be called widespread. An early use of regression weights was at Doane Agricultural Services Inc., now Doane Marketing Research. During 1971-1972 a readership study of farmers was conducted under the direction of Mr. John Wilkin in which 6,920 farmers responded. Weights for the respondents were constructed using regression procedures, where the controls came from the U.S. Agricultural Census and from Department of Agriculture sources. Doane provided financial support to Iowa State University to develop a regression weight generation program. To guarantee positive weights in the Doane study, observations with small weights were grouped and assigned a common weight. Grouping continued until the common weight was positive. Later computer programs used modifications of the Huang and Fuller (1978) procedure to guarantee positive weights. Doane has used regression weights for their syndicated market research studies since 1972.

Regression estimation was first used at Statistics Canada in 1988 for the Canadian Labour Force Survey. In 1992 regression estimation was used by the 1991 *Canadian Census of Population* to ensure that the weighted sum of variables collected via the long form (a one in five systematic sample of all households in Canada) was equal to known household and population totals as collected in the 1991 Census. See Bankier, Rathwell and Majkowski (1992) and Bankier, Houle and Luc (1997). The regression estimator is also the key component of the Generalized

Estimation System (GES) developed at Statistics Canada and used in numerous business and social surveys since its release in 1992. The methodology is described in Estevao, Hidirolou and Särndal (1995). See also Hidirolou, Särndal and Binder (1995). Regression estimation is now used to construct composite estimators for the Canadian Labour Force Survey. See Singh, Kennedy and Wu (2001), Gambino, Kennedy and Singh (2001) and Fuller and Rao (2001).

Bethlehem and Keller (1987) report on the use of regression estimation at the Netherlands Central Bureau of Statistics (now Statistics Netherlands) in a program called LIN WEIGHT. Nieuwenbroek, Renssen and Hofman (2000) describe the software package Bascula, that has replaced LIN WEIGHT. Deville, Särndal and Sautory (1993) describe a computer program CALMAR developed at Institut National de la Statistique et des Etudes Economiques (I. N. S. E. E.) that computes weights of the regression type with options for different objective functions. A program developed at Statistics Sweden and called CLAN97 is documented in Anderson and Nordberg (1998). Folsom and Singh (2000) discuss a procedure developed at the Research Triangle Institute.

3. THE CLASSICAL LINEAR MODEL

The classical linear model is the foundation for survey regression estimation, but the survey situation requires certain adaptations. To introduce regression estimation for survey samples, we review the classical linear model. Assume

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + e_i, \quad i = 1, 2, \dots, n, \\ e_i \sim \text{NI}(0, \sigma_e^2), \quad (3.1)$$

where e_i is independent of the k -dimensional row vectors \mathbf{x}_i for all i and j , and $\boldsymbol{\beta}$ is the unknown parameter column vector. We will also use matrix representations for the sample quantities. Thus, for a sample of n elements,

$$\mathbf{X}' = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n) \quad \text{and} \quad \mathbf{y}' = (y_1, y_2, \dots, y_n).$$

Given a sample of size n and treating the \mathbf{x}_i as fixed, the best (minimum mean squared error) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \sum_{i \in A} \mathbf{x}'_i y_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \quad (3.2)$$

where A is the set of indexes of the sample elements and we assume, as we will throughout, that the matrix to be inverted is nonsingular. If the e_i are not normally distributed, $\hat{\boldsymbol{\beta}}$ is the estimator with smallest variance in the class of linear unbiased estimators. The estimator of a linear combination of the coefficients, say $\theta_a = \sum_{j=1}^k \alpha_j \beta_j$, can be written as

$$\hat{\theta}_a = \sum_{i \in A} w_{ai} y_i$$

where the weights, w_{ai} , minimize the Lagrangean

$$\sum_{i \in A} w_{ai}^2 + \sum_{j=1}^k \lambda_j \left(\sum_{i \in A} w_{ai} x_{ij} - \alpha_j \right)$$

and the λ_j are Lagrange multipliers. The variance of $\hat{\theta}_a$ is

$$V\{\hat{\theta}_a\} = V\left\{ \sum_{i \in A} w_{ai} e_i \right\} = \sum_{i \in A} w_{ai}^2 \sigma_e^2$$

because the weights are functions of the \mathbf{x}_i and not of y_i .

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} V\left\{ \sum_{i \in A} \mathbf{b}'_i \right\} \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \\ = V\left\{ \sum_{i \in A} \mathbf{c}_i \right\} \quad (3.3)$$

where $\mathbf{b}'_i = \mathbf{x}'_i e_i$ and $\mathbf{c}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_i e_i$. Because e_i is independent of \mathbf{x}_j for all i and j ,

$$V\left\{ \sum_{i \in A} \mathbf{b}'_i \right\} = \sum_{i \in A} V\{\mathbf{b}'_i\} = \sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \sigma_e^2$$

and we obtain the familiar expression,

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \sigma_e^2.$$

The usual unbiased estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained by replacing σ_e^2 with the unbiased estimator of σ_e^2 obtained as the mean square of the residuals, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$. An estimator of the covariance matrix that estimates $V\{\sum_{i \in A} \mathbf{b}'_i\}$ directly is

$$\tilde{V}_b\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \sum_{i \in A} \hat{\mathbf{b}}'_i \hat{\mathbf{b}}_i \left(\sum_{i \in A} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \\ = \sum_{i \in A} \hat{\mathbf{c}}'_i \hat{\mathbf{c}}_i, \quad (3.4)$$

where $\hat{\mathbf{b}}'_i = \mathbf{x}'_i \hat{e}_i$ and $\hat{\mathbf{c}}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_i \hat{e}_i$. In the same way

$$\hat{V}_b\{\hat{\theta}_a\} = \sum_{i \in A} w_{ai}^2 \hat{e}_i^2 \quad (3.5)$$

is a linear combination of the elements of (3.4) and is a consistent estimator of $V\{\hat{\theta}_a\}$. The estimator (3.4) is a consistent estimator of $V\{\hat{\boldsymbol{\beta}}\}$ when the covariance matrix of the e_i is a diagonal matrix with bounded elements. Thus it is a more robust estimator. However, the estimator (3.4) is biased downward because the variance of \hat{e}_i is usually less than the variance of e_i . Two methods are available for reducing the bias. The first is to make a degrees-of-freedom adjustment by multiplying $\tilde{V}_b\{\hat{\boldsymbol{\beta}}\}$ by $(n-k)^{-1}n$, where k is the dimension of \mathbf{x}_i . An alternative adjustment is to replace \hat{e}_i with

$$\tilde{e}_i = (1 - \psi_{ii})^{-0.5} \hat{e}_i,$$

where ψ_{ii} is the i -th diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. See Horn, Horn and Duncan (1975), Royall and Cumberland (1978) and Cook and Weisberg (1982, section 2.2).

If we observe the value \mathbf{x}_i for an element, but do not observe y_i , then the best predictor of y_i for that element is $\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$. Likewise, if we know the sum of \mathbf{x}_i for a set of \mathbf{x} 's, then the best predictor for the sum of the y_i is the sum of $\mathbf{x}_i \hat{\boldsymbol{\beta}}$. Thus, given a set of N elements that satisfy model (3.1), a set of observations (y_i, \mathbf{x}_i) on a subset denoted by A , and the known values of \mathbf{x}_i for the remaining $N-n$ elements,

$$\hat{Y}_{N-n, \text{reg}} = \sum_{i \in \bar{A}} \hat{y}_i = \sum_{i \in \bar{A}} \mathbf{x}_i \hat{\boldsymbol{\beta}},$$

where \bar{A} is the set of elements for which y is not observed, is the best predictor of the sum of the unobserved y 's. See Goldberger (1962), Brewer (1963), Royall (1970), Harville (1976) and Graybill (1976, section 12.2). Hence

$$\hat{Y}_{y, \text{reg}} = \sum_{i \in A} y_i + \hat{Y}_{N-n, \text{reg}} \quad (3.6)$$

is the best predictor for the total of N observations.

If the first element in the x -vector is always one, we can partition the x -vector as $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$ and write the regression estimator of the mean as

$$\bar{y}_{\text{reg}} = N^{-1} \hat{Y}_{y, \text{reg}} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} = \bar{y}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n}) \hat{\boldsymbol{\beta}}_1, \quad (3.7)$$

where $\hat{\boldsymbol{\beta}}$ of (3.2) is partitioned as $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)'$ and $(\bar{y}_n, \bar{\mathbf{x}}_n)$ is the vector of simple sample means. We call $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ the regression estimator of the mean.

Given the model (3.1), the expected value of the mean of y for the finite population of N elements generated by the model is $\bar{\mathbf{x}}_N \boldsymbol{\beta}$ and $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ is an unbiased estimator of the finite population mean. This, we believe, is the point at which regression estimation for the finite population mean under more complex designs begins.

4. DESIGN BASED ESTIMATION

The development of this section treats the finite population as a sample realization from an infinite population. The use of such models has a long history in survey sampling. Some references through 1970 are Cochran (1939, 1942, 1946), Deming and Stephan (1941), Madow and Madow (1944), Yates (1949), Godambe (1955), Hájek (1959), Rao, Hartley, and Cochran (1962), Konijn (1962), Brewer (1963), Godambe and Joshi (1965), Hanurav (1966), Ericson (1969), Isaki (1970), and Royall (1970).

To discuss the large sample properties of regression estimators we consider sequences of finite populations and associated probability samples. The set of indices of the elements in the N th finite population is $U_N = \{1, \dots, N\}$, where $N = 1, 2, \dots$. Associated with the i th element of the N th population is a row vector of characteristics $\mathbf{z}_{iN} = (y_{iN}, \mathbf{x}_{iN})$. Let

$$\mathbf{F}_N = [(y_{1N}, \mathbf{x}_{1N}), (y_{2N}, \mathbf{x}_{2N}), \dots, (y_{NN}, \mathbf{x}_{NN})]$$

be the set of vectors for the N -th finite population. The subscript N on the vectors will often be omitted. The finite population mean is

$$\bar{\mathbf{z}}_N = (\bar{y}_N, \bar{\mathbf{x}}_N) = N^{-1} \sum_{i=1}^N (y_i, \mathbf{x}_i). \quad (4.1)$$

We denote the set of indices appearing in the sample selected from the N th finite population by A_N .

When the finite population is a sample from an infinite superpopulation, the probability properties of a sample are determined by the properties of the superpopulation and the properties of the probability mechanism used to select the sample. One can consider the unconditional properties, the properties conditional on the particular finite population, or the properties conditional on some part of the realized sample.

Properties conditional on the finite population depend primarily on the survey design and are often called design properties. Thus an estimator $\hat{\theta}$ is said to be design consistent for the finite population parameter θ_N if, for all $\epsilon > 0$,

$$\lim_{N, n \rightarrow \infty} \text{prob} \left\{ |\hat{\theta} - \theta_N| > \epsilon \mid \mathbf{F}_N \right\} = 0,$$

where the notation means that we condition on the realized finite population \mathbf{F}_N and, hence, the probability is with respect to the design.

Assume the finite population is generated as independent selections from a superpopulation for which $E\{\mathbf{z}_i' \mathbf{z}_i\}$ is positive definite, where $\mathbf{z}_i = (y_i, \mathbf{x}_i)$. We define a superpopulation vector of least squares regression coefficients by

$$\boldsymbol{\beta} = [E\{\mathbf{x}_i' \mathbf{x}_i\}]^{-1} E\{\mathbf{x}_i' y_i\}. \quad (4.2)$$

Given a sample of n observations on \mathbf{z}_i we define the $n \times (k+1)$ matrix $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$ of observations, where the i th row of \mathbf{Z} is (y_i, \mathbf{x}_i) . If we assume the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (4.3)$$

$$E\{\mathbf{u}, \mathbf{u}\mathbf{u}'\} = (\mathbf{0}, \boldsymbol{\Phi}),$$

the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}. \quad (4.4)$$

The model (4.3) serves as motivation for estimators of the form (4.4) but we shall consider estimators where $\boldsymbol{\Phi}$ is a general symmetric positive definite weight matrix, not necessarily the covariance matrix of the errors.

We give the large sample properties of the vector of estimated regression coefficients (4.4) following Fuller (1975). See also Hidirolou (1974), Scott and Wu (1981), and Robinson and Särndal (1983).

Assume the superpopulation has eighth moments and that the sample design is such that the error in the Horvitz-Thompson estimator of the mean is $O_p(n^{-1/2})$, where the Horvitz-Thompson estimator of the mean is

$$\bar{\mathbf{z}}_{\text{HT}} = (\bar{\mathbf{y}}_{\text{HT}}', \bar{\mathbf{x}}_{\text{HT}}') = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i \quad (4.5)$$

and π_i is the selection probability for element i . Then the error in the vector of regression coefficients is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N | \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1} \bar{\mathbf{b}}'_{\text{HT}} + O_p(n^{-1}), \quad (4.6)$$

where

$$\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1} \mathbf{Q}_{xyN}, \quad (4.7)$$

$$(\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}) = E\{(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) | \mathbf{F}_N\}, \quad (4.8)$$

$$(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy}) = n^{-1} (\mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{X}, \mathbf{X}' \boldsymbol{\Phi}^{-1} \mathbf{y}),$$

$$\bar{\mathbf{b}}_{\text{HT}} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i, \quad (4.9)$$

$\mathbf{b}_i' = n^{-1} N \pi_i \zeta_i' e_i$, $e_i = y_i - \mathbf{x}_i \boldsymbol{\beta}_N$, and ζ_i is column i of $\mathbf{X}' \boldsymbol{\Phi}^{-1}$. By (4.9) the error in the estimator of $\boldsymbol{\beta}_N$ is approximately the error in a Horvitz-Thompson estimator of the mean. In result (4.6), the $\boldsymbol{\beta}_N$ is defined as a function of the expected values of the sample quantities $(\hat{\mathbf{Q}}_{xx}, \hat{\mathbf{Q}}_{xy})$. Thus $\boldsymbol{\beta}_N$ is not necessarily the ordinary least squares finite population regression coefficient. The vector \mathbf{b}_i of (4.9) is the generalization of the vector \mathbf{b}_i of (3.3). If the limiting distribution of the properly standardized Horvitz-Thompson estimator is normal, and if there is a design consistent estimator of the variance of the Horvitz-Thompson estimator, then it is possible to construct tests and confidence intervals for the coefficients. Assume the design is such that

$$\mathbf{V}_{\bar{\bar{z}}}^{-1/2} (\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N) | \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad (4.10)$$

as $N, n \rightarrow \infty$, where $\mathbf{V}_{\bar{\bar{z}}}$ is the covariance matrix of $\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N$. If $\mathbf{V}_{\bar{\bar{z}}}$ is $O(n^{-1})$ and the estimator $\hat{\mathbf{V}}_{\bar{\bar{z}}}$ is consistent for $\mathbf{V}_{\bar{\bar{z}}}$, then

$$[\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\}]^{-1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) | \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad (4.11)$$

where

$$\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{V}}_{\bar{\bar{b}}} \hat{\mathbf{Q}}_{xx}^{-1} = \hat{\mathbf{V}}\{\bar{\mathbf{c}}'_{\text{HT}}\}, \quad (4.12)$$

$\hat{\mathbf{V}}_{\bar{\bar{b}}} = \hat{\mathbf{V}}\{\bar{\mathbf{b}}'_{\text{HT}}\}$ is the estimated design variance of $\bar{\mathbf{b}}_{\text{HT}}$ calculated with $\bar{\mathbf{b}}'_i = n^{-1} N \pi_i \zeta_i' \hat{e}_i$, $\hat{e}_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, and $\hat{\mathbf{V}}\{\bar{\mathbf{c}}'_{\text{HT}}\}$ is the estimated design variance of $\bar{\mathbf{c}}'_{\text{HT}}$ calculated with $\hat{\mathbf{c}}'_i = \hat{\mathbf{Q}}_{xx}^{-1} \bar{\mathbf{b}}'_i$. The limiting properties hold for stratified samples and for stratified two stage samples under mild restrictions on the sequence of populations.

By analogy to (3.7), a regression estimator of the finite population mean is obtained by evaluating the estimated regression function at the population mean of \mathbf{x} to obtain

$$\bar{\mathbf{y}}_{\text{reg}} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}, \quad (4.13)$$

where $\hat{\boldsymbol{\beta}}$ is of the form (4.4) with a general $\boldsymbol{\Phi}$ matrix. The estimator can be written as $\mathbf{w}' \mathbf{y}$, where the vector of weights can be constructed by minimizing the Lagrangean

$$\mathbf{w}' \boldsymbol{\Phi} \mathbf{w} + (\mathbf{w}' \mathbf{X} - \bar{\mathbf{x}}_N) \boldsymbol{\lambda}$$

and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers.

If there is a column vectors $\boldsymbol{\gamma}$ such that

$$\mathbf{X} \boldsymbol{\gamma} = \boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J} \quad (4.14)$$

for all possible samples, where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \dots, \pi_n)$ and \mathbf{J} is an n -dimensional column vector of ones, then the regression estimator $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ of (4.13) with $\hat{\boldsymbol{\beta}}$ defined in (4.4) is a design consistent estimator of $\bar{\mathbf{y}}_N$. It follows from (4.11) that

$$[\bar{\mathbf{x}}_N \hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\} \bar{\mathbf{x}}_N']^{-1/2} (\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} - \bar{\mathbf{y}}_N) \xrightarrow{L} N(0, 1). \quad (4.15)$$

The requirement of (4.14) that $\boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J}$ be in the column space of \mathbf{X} is crucial for design consistency. Simple ways to satisfy this requirement are to let one column of \mathbf{X} be the column of ones and to use a multiple of \mathbf{D}_π as $\boldsymbol{\Phi}$, or to let one column of \mathbf{X} be the elements π_i^{-1} and set $\boldsymbol{\Phi} = \mathbf{I}$, or to let one column of \mathbf{X} be the elements π_i and set $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$. If \mathbf{X} is composed of the single column vector with elements π_i and if $\boldsymbol{\Phi} = \mathbf{D}_\pi^2$, then the estimator (4.13) reduces to the Horvitz-Thompson estimator of (4.5) for fixed size designs. If $\mathbf{X} = \mathbf{J}$ and $\boldsymbol{\Phi} = \mathbf{D}_\pi$, the estimator (4.13) reduces to the ratio estimator,

$$\bar{\mathbf{y}}_\pi = \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} y_i, \quad (4.16)$$

which is location and scale invariant.

To see the nature of the estimator when (4.14) is satisfied, let, with no loss of generality, $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1)$, where $\mathbf{x}_0 = \boldsymbol{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J}$ and $\mathbf{x}_1 = (x_{0,i}, x_{1,i})$. Then

$$\bar{\mathbf{y}}_{\text{reg}} = \bar{\mathbf{x}}_{0,N} \bar{\mathbf{x}}_{0,\pi}^{-1} \bar{\mathbf{y}}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{0,N} \bar{\mathbf{x}}_{0,\pi}^{-1} \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_1, \quad (4.17)$$

where

$$\hat{\boldsymbol{\beta}}_1 = [(\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})' \boldsymbol{\Phi}^{-1} (\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})]^{-1} \times (\mathbf{X}_1 - \mathbf{x}_0 \hat{\boldsymbol{\mu}}_{x1})' \boldsymbol{\Phi}^{-1} \mathbf{y},$$

$\hat{\boldsymbol{\mu}}_{x1} = \bar{\mathbf{x}}_{0,\pi}^{-1} \bar{\mathbf{x}}_{1,\pi}$, and $(\bar{\mathbf{y}}_\pi, \bar{\mathbf{x}}_\pi)$ is defined in (4.16). The ratios, such as $\bar{\mathbf{x}}_{0,\pi}^{-1} \bar{\mathbf{y}}_\pi$, can also be written as ratios of Horvitz-Thompson estimators. If \mathbf{J} is in the column space of \mathbf{X} , estimator (4.17) is location invariant. If $\boldsymbol{\Phi} = \mathbf{D}_\pi$, then $\bar{\mathbf{x}}_{0,\pi}^{-1} \bar{\mathbf{x}}_{0,N} = 1$, and

$$\bar{\mathbf{y}}_{\text{reg}} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}} = \bar{\mathbf{y}}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_1, \quad (4.18)$$

where

$$\hat{\beta}_1 = \left[\sum_{i \in A} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})' \pi_i^{-1} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi}) \right]^{-1} \times \sum_{i \in A} (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi})' \pi_i^{-1} (y_i - \bar{y}_\pi). \quad (4.19)$$

Also, when $\Phi = \mathbf{D}_\pi$, the β_N of (4.7) is the population regression coefficient

$$\beta_N = \left[\sum_{i \in U} \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \sum_{i \in U} \mathbf{x}_i' y_i. \quad (4.20)$$

Because the regression estimator of the mean is a linear combination of regression coefficients, it is a regression coefficient for a linear combination of the original x -variables. To see this, let $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i})' = (1, \mathbf{x}_{1,i})'$, and define a new vector with one in the first position and a second vector with population mean equal to zero obtained by subtracting the original population mean $\bar{\mathbf{x}}_{1,N}$ from the original $\mathbf{x}_{1,i}$ vector. Let $\mathbf{q}_i = (1, \mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'$ be the transformed vector. Then the transformed regression model is

$$y_i = \mathbf{q}_i' \gamma + e_i, \quad (4.21)$$

where the finite population coefficient vector is

$$\gamma_N = (\bar{y}_N, \beta_{1,N})' = \left(\sum_{i \in U} \mathbf{q}_i' \mathbf{q}_i \right)^{-1} \sum_{i \in U} \mathbf{q}_i' y_i. \quad (4.22)$$

The expression for the regression estimator of the mean becomes

$$\bar{y}_{\text{reg}} = \bar{\mathbf{q}}_N' \hat{\gamma} = \hat{\gamma}_0, \quad (4.23)$$

where $\hat{\gamma}$ is obtained from (4.4) with \mathbf{q}_i replacing \mathbf{x}_i . Because the estimator is a linear estimator of the form $\mathbf{w}'\mathbf{y}$, we can write

$$\bar{y}_{\text{reg}} = \sum_{i \in A} w_i y_i = \sum_{i \in A} \pi_i^{-1} g_i y_i, \quad (4.24)$$

where $w_i = \pi_i^{-1} g_i$. Furthermore, the estimated variance from (4.12) is

$$\hat{V}\{\bar{y}_{\text{reg}}\} = \hat{V}\{\hat{\gamma}_0\} = \hat{V}\left\{ \sum_{i \in A} \pi_i^{-1} (g_i \hat{e}_i) \right\}, \quad (4.25)$$

where it is understood that the estimated design variance of (4.25) is computed for the variable $g_i \hat{e}_i$, $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\beta}$, and $\hat{\beta}$ is defined in (4.4). The variance estimator (4.25) is a direct generalization of expression (3.5). By transforming the variables so that the population mean of the auxiliary vector is zero, the first element of the regression vector is the regression estimator of the mean and the first element of (4.12) is an estimator of the variance of the regression estimator that contains a component due to estimating β . This was pointed out in Hidiroglou, Fuller, and Hickman (1978). Also, see Särndal (1982). Särndal, Swensson and Wretman (1989) suggested the g -factor terminology for the calculation of the estimated variance of a regression estimated total.

From (4.17), we can write

$$\begin{aligned} \bar{y}_{\text{reg}} &= \bar{x}_{0,N} \bar{x}_{0,\pi}^{-1} [\bar{y}_\pi - \bar{x}_{1,\pi} \beta_{1,N} - (\bar{y}_N - \bar{x}_{1,N} \beta_{1,N})] \\ &\quad + O_p(n^{-1}), \\ &= \bar{e}_\pi + O_p(n^{-1}), \end{aligned}$$

where $e_i = y_i - \mathbf{x}_i' \beta$. Hence, the variance of the regression estimator can be estimated with

$$\hat{V}\{\bar{e}_\pi\} = \hat{V}\left\{ \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} \hat{e}_i \right\}, \quad (4.26)$$

where $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\beta}$. Because (4.25) is as easy to compute as (4.26), and is applicable when $\bar{\mathbf{x}}_{1,\pi} - \bar{\mathbf{x}}_{1,N}$ is not $O_p(n^{-1/2})$, the estimator (4.25) is recommended.

The variance of the regression estimator can also be computed using the jackknife or other replication methods, and the use of replication methods is becoming more common. See Frankel (1971), Kish and Frankel (1974), Woodruff and Causey (1976), Royall and Cumberland (1978), and Duchesne (2000). Yung and Rao (1996) showed that (4.25) is identical to a jackknife linearization estimator for stratified multistage designs.

The approach to regression estimation associated with (4.18) and (4.19) falls completely within a design formulation. No models of the population, beyond the existence of moments, are used, through one might argue that one would only consider regression when one feels there is some linear correlation between $\mathbf{x}_{1,i}$ and y_i .

The estimator (4.19) is a very natural estimator because the estimated regression coefficient is a design consistent estimator of the population regression coefficient. It is mildly annoying that (4.18) does not always yield the smallest large sample design variance for the estimated mean. Treating $\hat{\beta}_1$ of (4.18) as a fixed vector, the value that minimizes the variance of the linear combination of means is

$$\beta_{1,\text{dopt}} = \left[V\{\bar{\mathbf{x}}_{1,\pi} | \mathbf{F}_N\} \right]^{-1} C\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi | \mathbf{F}_N\}. \quad (4.27)$$

See Cochran (1977, page 201), Fuller and Isaki (1981), Montanari (1987, 1999) and Rao (1994). If there is a design consistent estimator of the variance of $\bar{\mathbf{x}}_{1,\pi}$, then the $\beta_{1,d}$ that minimizes the estimated variance

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \beta_{1,d}\}, \quad (4.28)$$

denoted by $\hat{\beta}_{1,\text{dopt}}$, is a consistent estimator of $\beta_{1,\text{dopt}}$. It follows that the estimator

$$\bar{y}_{d,\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\beta}_{1,\text{dopt}} \quad (4.29)$$

has the minimum limit variance for design consistent estimators of the form $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \beta_{1,d}$. Also

$$\left[\hat{V}\{\bar{e}_\pi\} \right]^{-1/2} (\bar{y}_{d,\text{reg}} - \bar{y}_N) \xrightarrow{L} N(0, 1), \quad (4.30)$$

where $\hat{V}\{\bar{e}_\pi\}$ is the estimator of (4.26) constructed with $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \mathbf{x}_{1,\pi})\hat{\beta}_{1,\text{dopt}}$.

In a large sample sense, (4.29) answers the question of how to construct a regression estimator with optimum design properties. In practice a number of questions remain. The estimator is obtained under the assumption of a large sample and a vector \mathbf{x} of fixed dimension. In practice there may be a number of potential auxiliary variables and if a large number are included in the regression, terms excluded in the large sample approximation become important. This is particularly true for cluster samples where the number of primary sampling units in the sample is small. In such cases, the number of degrees-of-freedom in $\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\}$ is small and the inverse can be unstable. These issues are discussed further in section 9.

The estimator $\hat{\beta}_{1,\text{dopt}}$ of (4.29) is linear in y for most designs. See Rao (1994). For example, for a stratified design with simple random sampling within strata,

$$\hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\} = \sum_{h=1}^H K_h \sum_{j=1}^{n_h} (\mathbf{x}_{1,hj} - \bar{\mathbf{x}}_{1,h})' (y_{hj} - \bar{y}_h), \quad (4.31)$$

where

$$K_h = W_h^2 (1 - f_h) (n_h - 1)^{-1} n_h^{-1} = N^{-2} \pi_h^{-2} (1 - f_h) (n_h - 1)^{-1} n_h,$$

$N^{-1}N_h = W_h$, N_h is the size of stratum h , $f_h = \pi_h = N_h^{-1}n_h$, and n_h is the sample size in stratum h . It follows that the weights associated with estimator (4.29) are

$$w_{hi} = N^{-1} \pi_h^{-1} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \times \left[\sum_{t=1}^H K_t \sum_{j=1}^{n_h} (\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t})' (\mathbf{x}_{1,tj} - \bar{\mathbf{x}}_{1,t}) \right]^{-1} \times K_h (\mathbf{x}_{1,hi} - \bar{\mathbf{x}}_{1,h})'. \quad (4.32)$$

See also Särndal (1996). The weights of (4.32) can be constructed by minimizing $\sum_{hi \in A} w_{hi}^2 K_h^{-1}$ subject to the constraints

$$\sum_{i \in A_h} w_{hi} = N^{-1} N_h, \quad h = 1, 2, \dots, H,$$

and

$$\sum_{hi \in A} w_{hi} \mathbf{x}_{1,hi} = \bar{\mathbf{x}}_{1,N},$$

where A_h is the set of sample elements in stratum h .

The estimator of (4.19) with $\Phi = \mathbf{D}_\pi$ is a function of Horvitz-Thompson estimators of population moments. The estimator (4.17) with $\Phi^{-1} = \text{diag}\{K_t\}$, the diagonal matrix with K_t on the diagonal for elements in stratum t , and dummy variables for stratum effects, gives the estimator of the mean in the class

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\beta}_1$$

with the smallest estimated design variance. If the true slopes in the strata are the same and if the selection probabilities are proportional to the square roots of the within-stratum variances, then the use of $\Phi = \mathbf{D}_\pi^2$ gives a smaller small sample MSE than the use of $\Phi^{-1} = \text{diag}\{K_t\}$ because the sum of $w_{hi}^2 \sigma_h^2$ is smaller. Fuller and Isaki (1981) noted that the design-optimum estimator is often well approximated by the estimator constructed with $\Phi = \mathbf{D}_\pi^2$.

We have introduced regression estimation for the mean, but it is often the totals that are estimated and totals that are used as controls. Consider the regression estimator of the total of y defined by

$$\hat{T}_{y,\text{reg}} = \hat{T}_{y,\pi} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \hat{\beta}_{y,x}, \quad (4.33)$$

where $\mathbf{T}_{x,N}$ is the known total of \mathbf{x} and $(\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi})$ is a vector of design consistent estimators of $(T_{y,N}, \mathbf{T}_{x,N})$. By analogy to (4.28), the estimator of the optimum β is

$$\hat{\beta}_{y,x} = [\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}]^{-1} \hat{C}\{\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi}\}, \quad (4.34)$$

where $\hat{V}\{\hat{\mathbf{T}}_{x,\pi}\}$ is a design consistent estimator of the variance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{C}\{\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi}\}$ is a design consistent estimator of the covariance of $\hat{\mathbf{T}}_{x,\pi}$ and $\hat{T}_{y,\pi}$.

The estimator of the total is $N \bar{y}_{\text{reg}}$ for simple random sampling, but the exact equivalence may not hold in more complicated samples, because in such situations the estimated mean may be a ratio estimator. However, if the regression estimator of the two totals is constructed using (4.34), the ratio of the two estimated totals has large sample variance equal to that of the regression estimator of the mean. To see this write the error in the regression estimated totals of y and u as

$$\hat{T}_{y,\text{reg}} - T_{y,N} = \hat{T}_{y,\pi} - T_{y,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \beta_{y,x,N} + O_p(Nn^{-1})$$

and

$$\hat{T}_{u,\text{reg}} - T_{u,N} = \hat{T}_{u,\pi} - T_{u,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \beta_{u,x,N} + O_p(Nn^{-1}), \quad (4.35)$$

where we are assuming $\hat{T}_{y,\pi} - T_{y,N}$, $\hat{\beta}_{y,x} - \beta_{y,x,N}$ and the corresponding quantities for u , to be $O_p(Nn^{-1/2})$ and $O_p(n^{-1/2})$, respectively. Then the error in $\hat{T}_{u,\text{reg}}^{\beta_{y,x,N}^{-1}} \hat{T}_{y,\text{reg}}$ is

$$\begin{aligned} & \hat{T}_{u,\text{reg}}^{-1} \hat{T}_{y,\text{reg}} - T_{u,N}^{-1} T_{y,N} = T_{u,N}^{-1} \left[(\hat{T}_{y,\pi} - T_{y,N}) \right. \\ & \quad \left. - R_N (\hat{T}_{u,\pi} - T_{u,N}) \right] \\ & \quad + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) (\beta_{y,x,N} - R_N \beta_{u,x,N}) \\ & \quad + O_p(Nn^{-1}), \end{aligned} \quad (4.36)$$

where $R_N = T_{u,N}^{-1} T_{y,N}$. If we construct the regression estimator for R_N starting with $\hat{R} = \hat{T}_{u,\pi}^{-1} \hat{T}_{y,\pi}$, we have

$$\hat{R}_{\text{reg}} = \hat{R} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \hat{\boldsymbol{\beta}}_{R,x}, \quad (4.37)$$

where

$$\hat{\boldsymbol{\beta}}_{R,x} = [\hat{V}(\hat{\mathbf{T}}_{x,\pi})]^{-1} \hat{C}(\hat{\mathbf{T}}'_{x,\pi}, \hat{R})$$

and

$$\hat{C}(\hat{\mathbf{T}}_{x,\pi}, \hat{R}) = \hat{C}(\hat{\mathbf{T}}_{x,\pi}, T_{u,N}^{-1}(\hat{T}_{y,\pi} - R_N \hat{T}_{u,\pi})).$$

It follows that the large-sample-design-optimum coefficient for the ratio is $T_{u,N}^{-1}(\boldsymbol{\beta}_{y,x,N} - R_N \boldsymbol{\beta}_{u,x,N})$ and the ratio of design-optimum regression estimators is the large sample design-optimum regression estimator of the ratio.

5. MODELS AND REGRESSION ESTIMATION

In this section we assume that the analyst postulates a detailed superpopulation model. Assume also that the sample is an unequal probability sample or (and) the specified error covariance structure is not a multiple of the identity matrix. Then, only in special cases will the design optimal estimator of (4.29) agree with the best estimator constructed under the model, conditioning on the sample \mathbf{x} -values. To investigate this possible conflict, write the model for the population in matrix notation as

$$\begin{aligned} \mathbf{y}_U &= \mathbf{X}_U \boldsymbol{\beta} + \mathbf{e}_U \\ \mathbf{e}_U &\sim (\mathbf{0}, \boldsymbol{\Sigma}_{eeUU}), \end{aligned} \quad (5.1)$$

where $\mathbf{y}_U = (y_1, y_2, \dots, y_N)'$, $\mathbf{e}_U = (e_1, e_2, \dots, e_N)'$ and $\mathbf{X}_U = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_N)'$. It is assumed that $\boldsymbol{\Sigma}_{eeUU}$ is known or known up to a multiple. The model for a sample of n observations is

$$\begin{aligned} \mathbf{y}_A &= \mathbf{X}_A \boldsymbol{\beta} + \mathbf{e}_A, \\ \mathbf{e}_A &\sim (\mathbf{0}, \boldsymbol{\Sigma}_{eeAA}), \end{aligned}$$

where $\mathbf{y}_A = (y_1, y_2, \dots, y_n)'$, $\mathbf{e}_A = (e_1, e_2, \dots, e_n)'$, $\mathbf{X}_A = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n)'$, and we index the sample elements by 1, 2, ..., n , for convenience. We have used the subscript U to identify population quantities, and the subscript A to identify sample quantities, but we will often omit the subscript A to simplify the notation. For example, we may sometimes write the $n \times n$ covariance matrix as $\boldsymbol{\Sigma}_{ee}$. The unknown finite population mean is

$$\bar{y}_N = \bar{\mathbf{x}}_N \boldsymbol{\beta} + \bar{\mathbf{e}}_N. \quad (5.2)$$

Under model (5.1), the best linear, conditionally unbiased predictor of $\theta_N = \bar{y}_N$, conditional on \mathbf{X} is

$$\begin{aligned} \hat{\theta} &= N^{-1} \left[\sum_{i \in A} y_i + (N - n) \bar{\mathbf{x}}_{N-n} \hat{\boldsymbol{\beta}} \right. \\ &\quad \left. + \mathbf{J}'_{N-n} \boldsymbol{\Gamma}_{AA} (\mathbf{y}_A - \mathbf{X}_A \hat{\boldsymbol{\beta}}) \right], \end{aligned} \quad (5.3)$$

where $\boldsymbol{\Gamma}_{AA} = \boldsymbol{\Sigma}_{eeAA} \boldsymbol{\Sigma}_{eeAA}^{-1}$, $\bar{\mathbf{x}}_{N-n} = (N - n)^{-1} (N \bar{\mathbf{x}}_N - n \bar{\mathbf{x}}_n)$, $\boldsymbol{\Sigma}_{eeAA} = E\{\mathbf{e}_A \mathbf{e}_A'\}$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \boldsymbol{\Sigma}_{eeAA}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}_{eeAA}^{-1} \mathbf{y},$$

$\mathbf{e}_A = (e_{n+1}, e_{n+2}, \dots, e_N)$, \mathbf{J}_{N-n} is an $N - n$ dimensional column vector of ones, $\bar{\mathbf{x}}_n$ is the simple sample mean, and A is the set of elements in U that are not in A . See Royall (1976). Under the model,

$$\hat{\theta} - \bar{y}_N = \mathbf{C}_{x\bar{A}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + N^{-1} \mathbf{J}'_{N-n} (\boldsymbol{\Gamma}_{AA} \mathbf{e}_A - \mathbf{e}_{\bar{A}})$$

and

$$\begin{aligned} V\{\hat{\theta} - \bar{y}_N | \mathbf{X}_A\} &= \mathbf{C}_{x\bar{A}} V\{\hat{\boldsymbol{\beta}}\} \mathbf{C}_{x\bar{A}}' \\ &\quad + N^{-2} \mathbf{J}'_{N-n} (\boldsymbol{\Sigma}_{ee\bar{A}\bar{A}} - \boldsymbol{\Gamma}_{AA} \boldsymbol{\Sigma}_{eeAA}) \mathbf{J}_{N-n}, \end{aligned} \quad (5.4)$$

where

$$\mathbf{C}_{x\bar{A}} = N^{-1} [(N - n) \bar{\mathbf{x}}_{N-n} - \mathbf{J}'_{N-n} \boldsymbol{\Gamma}_{AA} \mathbf{X}_A].$$

Design consistency of estimator (5.3) and the situations in which the model estimator reduces to the Horvitz-Thompson estimator have been considered by, among others, Isaki (1970), Royall (1970, 1976), Scott and Smith (1974), Cassel, Särndal, and Wretman (1976, 1979, 1983), Zyskind (1976), Tallis (1978), Isaki and Fuller (1982), Wright (1983), Pfefferman (1984), Tam (1986), Brewer, Hanif and Tam (1988), Montanari (1999), and Gerow and McCulloch (2000).

The estimator (5.3) reduces to $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ if there is an η such that

$$\mathbf{X}_A \eta = \boldsymbol{\Sigma}_{eeAA} \mathbf{J}_n + \boldsymbol{\Sigma}_{eeA\bar{A}} \mathbf{J}_{N-n}, \quad (5.5)$$

for all samples with positive probability. If there is also γ such that

$$\mathbf{X}_A \gamma = \boldsymbol{\Sigma}_{eeAA} \mathbf{D}_\pi^{-1} \mathbf{J}_n \quad (5.6)$$

for all samples with positive probability, then $\hat{\theta}$ of (5.3) is design consistent, where \mathbf{D}_π was defined for (4.14). Given a \mathbf{k} such that

$$\mathbf{X}_A \mathbf{k} = \boldsymbol{\Sigma}_{eeAA} (\mathbf{D}_\pi^{-1} \mathbf{J}_n - \mathbf{J}_n) - \boldsymbol{\Sigma}_{eeA\bar{A}} \mathbf{J}_{N-n}, \quad (5.7)$$

then $\hat{\theta}$ of (5.3) is expressible as

$$\hat{\theta} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi) \hat{\boldsymbol{\beta}} \quad (5.8)$$

and if the design is such that \bar{y}_π is design consistent for \bar{y}_N , $\hat{\theta}$ of (5.8) is design consistent for \bar{y}_N .

We call a regression model of the form (5.1) for which (5.5) and (5.6), or (5.7), holds a full model. If (5.6) or (5.7) does not hold, we call the model a reduced model or a restricted model. We cannot expect the conditions for a full model to hold for every analysis variable in a general purpose survey because $\boldsymbol{\Sigma}_{ee}$ will be different for different

y 's. Therefore, given a reduced model, one might search for a good model estimator in the class of design consistent estimators.

To construct a design consistent estimator of the form $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ when model (5.1) is a reduced model, we can add a vector satisfying (5.7) to the \mathbf{X} -matrix to create a full model. There are two possible situations associated with this approach. In the first, the population mean (or total) of the added variable is known. With known mean, one can construct the usual regression estimator and the usual design variance estimation formulas are appropriate.

To describe an estimation procedure for the situation in which the population mean of the added variable is not known, let $\mathbf{q} = (q_1, q_2, \dots, q_n)'$ denote the added vector, where \mathbf{q} is the vector on the right side of the equality in (5.7). Let $\mathbf{H} = (\mathbf{X}, \mathbf{q})$, where \mathbf{X} is the matrix of auxiliary variables with known population mean vector, $\bar{\mathbf{x}}_N$. We write the full model for the sample as

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta}_{y,h} + \mathbf{e}, \quad (5.9)$$

where $\mathbf{e} \sim (0, \boldsymbol{\Sigma}_{ee})$. The best linear conditionally unbiased estimator of $\boldsymbol{\beta}_{y,h}$ is

$$\hat{\boldsymbol{\beta}}_{y,h} = (\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{H})^{-1}\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{y}. \quad (5.10)$$

If the coefficient for \mathbf{q} in (5.9) is not zero, it is not possible to construct a conditionally unbiased estimator of $\mathbf{h}_N \boldsymbol{\beta}_{y,h}$ because the \bar{q}_N component of $\bar{\mathbf{h}}_N$ is unknown. However, because $\hat{\boldsymbol{\beta}}_{y,h}$ is unbiased for $\boldsymbol{\beta}_{y,h}$, it is possible to construct a conditionally unbiased estimator of any linear function of $\boldsymbol{\beta}_{y,h}$. Thus, it is natural to replace the unknown \bar{q}_N with the "best available" estimator of \bar{q}_N , and a reasonable choice is the regression estimator,

$$\bar{q}_{reg} = \bar{q}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi) \hat{\boldsymbol{\beta}}_{q,x}, \quad (5.11)$$

where $\hat{\boldsymbol{\beta}}_{q,x} = (\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{q}$. Then the estimator (5.3) becomes

$$\hat{\theta} = \bar{y}_\pi + [(\bar{\mathbf{x}}_N, \bar{q}_{reg}) - (\bar{\mathbf{x}}_\pi, \bar{q}_\pi)] \hat{\boldsymbol{\beta}}_{y,h} \quad (5.12)$$

The estimator (5.12) can be expressed in the familiar regression estimator form,

$$\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi) \hat{\boldsymbol{\beta}}_{y,x}. \quad (5.13)$$

That is, the regression estimator of the finite population mean of y based on the full model, but with the mean of q_i unknown and estimated with the regression estimator, is the regression estimator with $\hat{\boldsymbol{\beta}}_{y,x}$ estimated by the generalized least squares regression of y on \mathbf{x} using the covariance matrix $\boldsymbol{\Sigma}_{ee}$. See Park (2002). The estimator is conditionally model unbiased under the reduced model containing only \mathbf{x} if the reduced model is true. If the population coefficient for q_i is not zero, the reduced model is not true. Then the estimator is conditionally model biased, but the estimator is unbiased for the finite population mean under the full model and an unbiased design, because

$$\begin{aligned} E\{\bar{y}_{reg} - \bar{y}_N\} &= E\{E[\bar{y}_{reg} - \bar{y}_N | \mathbf{H}]\} \\ &= E\{((0, \bar{q}_{reg} - \bar{q}_N)\boldsymbol{\beta}_{y,h})\} = 0, \end{aligned} \quad (5.14)$$

where \bar{y}_{reg} is defined in (5.12) and the approximation is due to the approximate design expectation of the regression estimator \bar{q}_{reg} .

The estimator (5.13) is a linear estimator, where the vector of weights, \mathbf{w} , minimizes the Lagrangean

$$\mathbf{w}'\boldsymbol{\Sigma}_{ee}\mathbf{w} + [\mathbf{w}'\mathbf{H} - (\bar{\mathbf{x}}_N, \bar{q}_{reg})]\boldsymbol{\lambda}. \quad (5.15)$$

The estimator is location invariant if the column of ones is in the column space of \mathbf{X} .

Because the variable q is the variable whose omission from the full model can produce a bias, it seems prudent to test the coefficient of q before using the reduced model to construct an estimator for the mean of y . This can be done using a model estimator of the variance,

$$\hat{V}\{\hat{\boldsymbol{\beta}}_{y,h} | \mathbf{H}\} = (\mathbf{H}'\boldsymbol{\Sigma}_{ee}^{-1}\mathbf{H})^{-1}$$

or using the design estimator of variance of (4.12). See Du Mouchel and Duncan (1983) and Fuller (1984).

A working specification for $\boldsymbol{\Sigma}_{ee}$ may be particularly appropriate for two-stage samples, see Royall (1976, 1986) and Montanari (1987). A reasonable model is that in which there is common correlation among items in the same primary sampling unit and zero correlation between units in different primary sampling units. Because the associated $\boldsymbol{\Sigma}_{ee}$ is block diagonal of a particular form, it is relatively easy to invert and hence the estimator based on such a working $\boldsymbol{\Phi}$ is relatively easy to construct. The regression estimator using a $\boldsymbol{\Phi}$ with a non zero correlation for units in the same primary sampling unit is a combination of the estimator based on primary sampling unit totals and that based on elements. See Fuller and Battese (1973). Thus, the use of such a $\boldsymbol{\Phi}$ can avoid variance problems associated with the use of primary sampling unit totals.

6. MAXIMUM LIKELIHOOD AND RAKING RATIO

The theoretical foundation for the regression estimators discussed in section 3 and section 4 is maximum likelihood estimation for the linear model with normal errors. We now consider the likelihood for multinomial variables. Given a simple random sample from a multinomial defined by the entries in a two way table, the logarithm of the likelihood, except for a constant, is

$$\sum_{i=1}^r \sum_{j=1}^c a_{ij} \log p_{ij}, \quad (6.1)$$

where a_{ij} is the estimated fraction in cell ij , p_{ij} is the population fraction in cell ij , r is the number of rows, and c is the number of columns. If (6.1) is maximized subject to the restriction $\sum \sum p_{ij} = 1$, one obtains the maximum

likelihood estimators $\hat{p}_{ij} = a_{ij}$. If the marginal row fractions $p_{i\cdot,N}$ and the marginal column fractions $p_{\cdot j,N}$ are known, it is natural to maximize the likelihood subject to these constraints by using the Lagrangean

$$\sum_{i=1}^r \sum_{j=1}^c a_{ij} \log p_{ij} + \sum_{i=1}^r \lambda_i \left(\sum_{j=1}^c p_{ij} - p_{i\cdot,N} \right) + \sum_{j=r+1}^{r+c} \lambda_j \left(\sum_{i=1}^r p_{ij} - p_{\cdot j,N} \right), \quad (6.2)$$

where $\lambda_i, i = 1, 2, \dots, r$, are for the row restrictions and $\lambda_j, j = 1, 2, \dots, c$, are for the column restrictions. There is no explicit expression for the solution to (6.2) and there may be no solution if there are too many empty cells. A procedure that produces estimates close to the maximum likelihood solution is that called *raking ratio* or *iterative proportional fitting*. The procedure iterates, first making ratio adjustments for the row restrictions, then making ratio adjustments for the column restrictions, then making a ratio adjustments for the row restrictions, etc. The method is generally credited to Deming and Stephan (1940). See, for example, Bishop, Fienberg and Holland (1975, Chapter 3).

Deville and Särndal (1992) considered a class of objective functions of the form $\sum_{i \in A} G(w_i, \alpha_i)$, where $G(w, \alpha)$ is a measure of distance between an initial weight α_i and a final weight w_i . The objective function is minimized subject to the constraints

$$\sum_{i \in A} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N. \quad (6.3)$$

Deville and Särndal (1992) used the term *calibrated* to describe weights satisfying (6.3). If the initial weight is $\alpha_i = (\sum \pi_j^{-1})^{-1} \pi_i^{-1}$ and if one is the first element of \mathbf{x}_i , the solution to the minimization problem is approximated by a regression estimator of the mean of the form

$$\bar{y}_{\text{reg}} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)' \hat{\boldsymbol{\beta}}, \quad (6.4)$$

where

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i \in A} \mathbf{x}_i' \Phi_{ii}^{-1} \mathbf{x}_i \right]^{-1} \sum_{i \in A} \mathbf{x}_i' \Phi_{ii}^{-1} y_i,$$

and Φ_{ii} is the second derivative of $G(w, \alpha)$ with respect to w evaluated at $(w, \alpha) = (\alpha_i, \alpha_i)$. Using this approach, Deville and Särndal (1992) showed that the maximum likelihood and raking ratio estimators have the same limiting distribution as the regression estimator (4.18) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$. To obtain the raking ratio weights they used the objective function

$$\sum_{i \in A} \left[w_i \log \alpha_i^{-1} w_i + \alpha_i - w_i \right], \quad (6.5)$$

and to obtain the maximum likelihood weights they used the objective function

$$\sum_{i \in A} \left[w_i - \alpha_i - \alpha_i \log \alpha_i^{-1} w_i \right]. \quad (6.6)$$

Deville, Särndal and Sautory (1993) investigated four estimators in the class. Although weights constructed using different functions could differ considerably, the authors concluded that estimates were quite similar, a result consistent with the theory. Singh and Mohl (1996) and Th  berge (1999, 2000) discuss estimators with the calibration property.

7. POPULATION OF AUXILIARY VECTORS KNOWN AT ESTIMATION STEP

If the \mathbf{x} -vector is known for all of the population elements, the number of possible regression-type estimators is greatly expanded. Most procedures involve the fitting of an approximating function for the relationship between y and the auxiliary variables. The most used procedure is to assign the population elements to categories on the basis of the auxiliary data and to use these categories as post strata. This procedure is equivalent to approximating the expected value of y given \mathbf{x} by a step function. The estimator is formally equivalent to the regression estimator (4.19) where the \mathbf{x} -vector is a vector of indicator variables for post-stratum membership.

The application of the procedure often requires the development of criteria to use in forming the post strata. Typically the post strata are formed so that each post stratum contains a minimum number of sample elements and so that the weights for any post stratum are not overly large. Estimation with post strata and the formation of post strata have been studied by Fuller (1966), Holt and Smith (1979), Tremblay (1986), Kalton and Maligalig (1991), Little (1993), Eltinge and Yansaneh (1997), and Lazzeroni and Little (1998), among others. Holt and Smith (1979) argued for the use of a conditional variance estimator for post stratification.

Given the population of \mathbf{x} -vectors, one can use the sample to estimate a functional relationship between y and \mathbf{x} and then predict the unobserved y . If the procedure is to be design consistent, then a condition similar to (4.14) must hold. One way to ensure design consistency is to require the fitted model to satisfy

$$\sum_{i \in A} \pi_i^{-1} [y_i - f(\mathbf{x}_i; \hat{\boldsymbol{\beta}})] = 0, \quad (7.1)$$

where $f(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ is the model estimated value for the i -th observation.

Firth and Bennett (1998) pointed out that some nonlinear models satisfy (7.1). If the initial model does not satisfy (7.1), an estimated intercept term can be added to create an expanded full model,

$$\begin{aligned} \tilde{f}_F(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) &= f(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) \\ &+ \left(\sum_{i \in A} \pi_i^{-1} \right)^{-1} \sum_{i \in A} \pi_i^{-1} [y_i - f(\mathbf{x}_i; \hat{\boldsymbol{\beta}})]. \end{aligned}$$

This is a direct extension of the ideas of difference estimation to the nonlinear case. See Isaki (1970), Cassel, Särndal and Wretman (1976) and Wright (1983). A closely related approach was suggested by Wu and Sitter (2001) in which the fitted function $f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ is used as the auxiliary variable in a linear regression estimator.

A number of “local” procedures, other than step functions, can be used to approximate the functional relationship between \mathbf{x} and y . Spline functions and polynomials are linear models that fall within the class of section 4. Estimators that use some kind of local smoothing to estimate population quantities have been considered for finite populations from a model viewpoint by Kuo (1988), Dorfman (1993), Dorfman and Hall (1993), Chambers (1996), and Chambers, Dorfman and Wehrly (1993). Breidt and Opsomer (2000) showed that estimators based on local polynomial regression are design consistent. Firth and Bennett (1998) also considered local fit models.

8. REGRESSION ESTIMATION AND NONRESPONSE

Regression estimation is frequently a part of procedures used to adjust data for unit nonresponse. Regression can be justified on the basis of a model such as (3.1) or on the basis that regression can adjust for unequal response probabilities. See Cassel, Särndal and Wretman (1979, 1983), Little (1982, 1986), Bethlehem (1988), Kott (1994), Fuller, Loughin and Baker (1994) and Fuller and An (1998).

Consider an estimator of the population regression vector of the form (4.4) with $\boldsymbol{\Phi} = \mathbf{D}_\pi$ constructed with the responding units. Denote the estimator by $\hat{\boldsymbol{\beta}}$ and let p_i be the conditional probability of observing unit i given that the unit is selected for the sample. Then under regularity conditions, the estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of

$$\boldsymbol{\gamma}_N = \left(\sum_{i \in U} \mathbf{x}_i' p_i \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' p_i y_i. \quad (8.1)$$

The population mean of y can be expressed as

$$\bar{y}_N = \bar{\mathbf{x}}_N \boldsymbol{\gamma}_N + \bar{a}_N \quad (8.2)$$

where $a_i = y_i - \mathbf{x}_i' \boldsymbol{\gamma}_N$ and \bar{a}_N is the population mean of the a_i . The regression estimator $\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ will be consistent for \bar{y}_N if the probability limit of \bar{a}_N is zero. The probability limit of \bar{a}_N will be zero if the sequence of finite populations is a sequence of random samples from an infinite population in which

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \quad (8.3)$$

and the e_i of the sample are independent of \mathbf{x}_i with $E\{e_i | \mathbf{x}_i\} = 0$.

Alternatively, a sufficient condition for \bar{a}_N to be zero is the existence of a column vector $\boldsymbol{\xi}$ such that

$$\mathbf{x}_i \boldsymbol{\xi} = p_i^{-1} \quad (8.4)$$

for $i = 1, 2, \dots, N$. Thus, if the reciprocal of the response probability is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y . One way in which (8.4) can be satisfied is for the elements of \mathbf{x}_i to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup.

If (8.4) holds and if the probability of responding is independent from unit to unit, then the estimated variance based on (4.12) is an appropriate estimator for the variance of the regression estimator of the mean. It is particularly important that a variance estimator of the form (4.12) or (4.25), and not of the form (4.26) be used, because $\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi$ is, in general, not $O_p(n^{-1/2})$ in the presence of nonresponse. Singh and Folsom (2000) make a similar argument for the variance estimator (4.25) when using regression to adjust for coverage error.

Often a preliminary adjustment to the selection probabilities is made for nonresponse and this is followed by regression estimation. The most frequently used response adjustment is to form adjustment cells (post strata) and to ratio adjust the weights of respondents in the cell so that the sum of the weights is equal to the estimated (or known) total for the cell. See, for example, Little and Rubin (1987, page 250). Procedures using an estimated response probability function are discussed by Cassel, Särndal and Wretman (1983), Rosenbaum and Rubin (1983), Folsom and Witt (1994). Fuller and An (1998), and Folsom and Singh (2000). Brick, Waksberg and Keeter (1996) use an estimated contact probability to adjust for frame coverage.

To consider procedures based on estimated response probabilities, assume that the inverse of the response probability for individual i is given by

$$p_i^{-1} = g(\mathbf{z}_i; \boldsymbol{\theta}^0), \quad (8.5)$$

where \mathbf{z}_i is a vector of variables that can be observed for both respondents and nonrespondents, $\boldsymbol{\theta}^0$ is the true value of $\boldsymbol{\theta}$, and $g(\mathbf{z}_i; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ with continuous first and second derivatives in an open set containing $\boldsymbol{\theta}^0$ for all \mathbf{z}_i . The vector $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ is observed, and we assume that p_i is bounded below by a positive number.

Let δ_i be the indicator variable with $\delta_i = 1$ if a response is obtained and $\delta_i = 0$ if a response is not obtained. Using the vector (δ_i, \mathbf{z}_i) , the parameter $\boldsymbol{\theta}^0$ of the response probability function is estimated. Assume that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O_p(n^{-1/2})$, where $\hat{\boldsymbol{\theta}}$ is the estimator of $\boldsymbol{\theta}$. Let $\boldsymbol{\beta}_N$ denote the finite population regression vector for the regression of y on \mathbf{x} . Let

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{x}_i' \pi_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' y_i \pi_i^{-1} \hat{p}_i^{-1} \delta_i, \quad (8.6)$$

where π_i are the selection probabilities and $\hat{p}_i^{-1} = g(\mathbf{z}_i; \hat{\boldsymbol{\theta}})$. Under conditions of the type used in section 4,

$$\begin{aligned} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= \mathbf{M}_{xx}^{-1} \sum_{i \in A} \delta_i \pi_i^{-1} p_i^{-1} \mathbf{x}_i' a_i \left[1 + p_i \mathbf{g}_{1,i}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \right] \\ &\quad + O_p(n^{-1}), \end{aligned}$$

where $\mathbf{g}_{1,i}$ is the row vector of first derivatives of $g(\mathbf{z}_i; \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ and $\mathbf{M}_{xx} = \sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} p_i^{-1} \delta_i$. If $\mathbf{g}_{1,i}$ is uncorrelated with a_i , then the term involving $\mathbf{g}_{1,i} a_i$ is $O_p(n^{-1})$ and the variance estimator constructed as if $g(\mathbf{z}; \boldsymbol{\theta}^0)$ is known is appropriate. The conditions are satisfied if \mathbf{z}_i is a subvector of \mathbf{x}_i and \mathbf{z}_i defines imputation cells (adjustment cells) with equal response rates within a cell.

9. PRACTICAL CONSIDERATIONS

If the regression weights are to be used in a general purpose survey, no individual weight used in estimating a total should be less than one. Also, it seems reasonable, on robustness grounds, to avoid very large weights. We discuss some procedures that have been developed to accomplish these objectives.

A number of algorithms produce positive weights with a high probability. Raking ratio procedures produces positive weights for most data configurations. Deville, Särndal and Sautory (1993) discuss the extension of raking ratio to general x -variables and extensions to include bounds on the weights.

Tillé (1998) suggested the use of approximate conditional probabilities, conditional on $\bar{\mathbf{x}}_\pi$, to compute an estimator. His approximation can be extended to produce regression weights that are positive with high probability. Let $\bar{\mathbf{x}}_\pi^{(i)}$ be an estimator obtained by deleting element i , or primary sampling unit i , and modifying the remaining weights so that $\bar{\mathbf{x}}_\pi^{(i)}$ is unbiased, or consistent to the same order as $\bar{\mathbf{x}}_\pi$, for the population mean of all elements excluding i . The estimator $\bar{\mathbf{x}}_\pi^{(i)}$ can be the estimator used to construct jackknife deviates. Let $\hat{\Sigma}_{xx}$ be an estimator of the covariance matrix of $\bar{\mathbf{x}}_\pi$ and let $\hat{\Sigma}_{xx(i)}$ be an estimator of the conditional covariance matrix of $\bar{\mathbf{x}}_\pi$ conditional on $i \in A$. Then, in large samples $\bar{\mathbf{x}}_\pi$ and $\bar{\mathbf{x}}_\pi^{(i)}$ are approximately normally distributed and an estimator of the probability that i is in the sample given the estimated mean $\bar{\mathbf{x}}_\pi$, is

$$\begin{aligned} \hat{\pi}_{i|A} &= \hat{P}\{i \in A \mid \mathbf{F}_N, \bar{\mathbf{x}}_\pi\} \\ &= \pi_i \left| \frac{\hat{\Sigma}_{xx}^{-1/2}}{\hat{\Sigma}_{xx(i)}^{-1/2}} \right|^{-1/2} \exp \left\{ 0.5 (\mathbf{G}_{xx} - \mathbf{G}_{xx(i)})' \right\} \quad (9.1) \end{aligned}$$

where

$$\begin{aligned} \mathbf{G}_{xx} &= (\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N) \hat{\Sigma}_{xx}^{-1} (\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N)', \\ \mathbf{G}_{xx(i)} &= (\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N) \hat{\Sigma}_{xx(i)}^{-1} (\bar{\mathbf{x}}_\pi^{(i)} - \bar{\mathbf{x}}_N)', \end{aligned}$$

and $\bar{\mathbf{x}}_N^{(i)} = (N-1)^{-1} (N \bar{\mathbf{x}}_N - \mathbf{x}_i)$. For simple random sampling, Tillé (1998) showed that the estimator

$$\bar{y}_{p\pi} = N^{-1} \sum_{i \in A} \pi_i^{-1} y_i, \quad (9.2)$$

where $\pi_{i|A}$ is the conditional probability calculated under the normality assumptions, is approximately equal to the

regression estimator. Because the estimator is not calibrated, we suggest a calibrated version obtained by computing the regression estimator with $\hat{\pi}_{i|A}$ as initial weights. The difference between (9.2) and the regression estimator constructed with initial weights $\hat{\pi}_{i|A}$ is $O_p(n^{-1})$. Hence, there is a good chance that the regression weights so constructed will be positive. The variance estimator $\hat{\Sigma}_{xx(i)}$ is relatively simple to compute for stratified samples but may require considerable computation for other cases. Thus one may choose to approximate $\Sigma_{xx(i)}$.

Given that the regression weights are being constructed by minimizing an objective function, one can add restrictions to the problem to place bounds on the weights. Huang and Fuller (1978) gave an iterative procedure equivalent to constructing a Φ matrix at each step that reduces the weight on observations whose current weight deviates from the average by a large absolute amount.

To discuss additional procedures associated with quadratic objective functions, assume we have a working covariance matrix, denoted by Φ_{ee} , for the model (5.1) that is to be used to construct a regression estimator. Let \mathbf{a} be the column vector of initial weights and assume $\Phi_{ee} \mathbf{a}$ is in the column space of \mathbf{X} . Then the weights that minimize the conditional model variance are the weights that minimize $\mathbf{w}' \Phi_{ee} \mathbf{w}$ or, equivalently, that minimize

$$(\mathbf{w} - \mathbf{a})' \Phi_{ee} (\mathbf{w} - \mathbf{a}) \quad (9.3)$$

subject to the constraint

$$\mathbf{w}' \mathbf{X} = \bar{\mathbf{x}}_N. \quad (9.4)$$

Given an objective function, we can add restrictions on the w_i such as

$$L_1 \leq w_i \leq L_2, \quad i \in A, \quad (9.5)$$

where L_1 and L_2 are nonnegative constants. Minimizing (9.3), subject to the constraints (9.4) and (9.5) is a quadratic programming problem. The use of quadratic programming was suggested by Husain (1969) and was used by Isaki, Tsay and Fuller (2000).

If a large number of control variables are used, it may not be possible to construct weights satisfying the calibration constraints and also falling within reasonable bounds. The practitioner is faced with making compromises. The most common practice is to drop variables from the model. See Bankier, Rathwell and Majkowski (1992) and Silva and Skinner (1997). To discuss an alternative procedure, consider the situation in which some of the constraints are required but others can be relaxed. Let the matrix of observations on the auxiliary variables be partitioned as $(\mathbf{X}_0, \mathbf{X}_2)$, where \mathbf{X}_0 is the set of variables for which exact constraints are required and \mathbf{X}_2 is the set for which the constraints can be relaxed. Assume $\Phi_{ee} \mathbf{a}$ is in the column space of \mathbf{X}_0 . Then a generalization of (9.3) and (9.4) is the function

$$(\mathbf{w} - \mathbf{a})' \Phi_{ee} (\mathbf{w} - \mathbf{a}) + (\mathbf{w}' \mathbf{X}_2 - \bar{\mathbf{x}}_{2,N}) \Psi (\mathbf{w}' \mathbf{X}_2 - \bar{\mathbf{x}}_{2,N})' \quad (9.6)$$

and the constraint

$$\mathbf{w}'\mathbf{X}_0 - \bar{\mathbf{x}}_{0,N} = \mathbf{0}, \quad (9.7)$$

where Φ_{ee} and Ψ are positive definite symmetric matrices and $\bar{\mathbf{x}}_N = (\bar{\mathbf{x}}_{0,N}, \bar{\mathbf{x}}_{2,N})$. The \mathbf{w} that minimizes (9.6) subject to (9.7) minimizes the mean squared error of the unbiased linear predictor of $\bar{\mathbf{x}}_N\boldsymbol{\beta}$ under the mixed model

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e},$$

where $\boldsymbol{\beta}_2 \sim (0, \Psi)$, $\mathbf{e} \sim (0, \Phi_{ee})$, the random vector $\boldsymbol{\beta}_2$ is independent of \mathbf{e} , and $\boldsymbol{\beta}_0$ is a fixed vector. See Lazzeroni and Little (1998) for the use of random models for post stratification.

The vector \mathbf{w}' that minimizes (9.6) subject to restriction (9.7) is

$$\mathbf{w}' = \boldsymbol{\alpha}' + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_N) \mathbf{H}_{x\psi x}^{-1} \mathbf{X}' \Phi_{ee}^{-1}, \quad (9.8)$$

where

$$\mathbf{H}_{x\psi x} = \begin{pmatrix} \mathbf{X}_0' \Phi_{ee}^{-1} \mathbf{X}_0 & \mathbf{X}_0' \Phi_{ee}^{-1} \mathbf{X}_2 \\ \mathbf{X}_2' \Phi_{ee}^{-1} \mathbf{X}_0 & \Psi^{-1} + \mathbf{X}_2' \Phi_{ee}^{-1} \mathbf{X}_2 \end{pmatrix}. \quad (9.9)$$

The estimator can be written

$$\bar{y}_{\text{reg}} = \mathbf{w}'\mathbf{y} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_N) \hat{\boldsymbol{\theta}}, \quad (9.10)$$

where $\hat{\boldsymbol{\theta}} = \mathbf{H}_{x\psi y}^{-1} \mathbf{X}' \Phi_{ee}^{-1} \mathbf{y}$. See Henderson (1963), Robinson (1991), and Rao (2002, Chapter 6).

Husain (1969) considered (9.6) for a simple random sample from a normal distribution with $\mathbf{X}_0 = \mathbf{J}$, $\Phi_{ee} = \mathbf{I}$, and $\Psi^{-1} = \gamma^{-1} \hat{\Sigma}_{x,22}$, where $\hat{\Sigma}_{x,22}$ is the estimated covariance matrix of $\bar{\mathbf{x}}_{2,\pi}$, and γ is a constant to be determined. For this case, Husain showed that the optimal γ is

$$\gamma_{\text{opt}} = [k_2(1 - R^2)]^{-1}(n - k_2 - 2)R^2, \quad (9.11)$$

where k_2 is the dimension of \mathbf{x}_2 and R^2 is the squared multiple correlation coefficient. Bardsley and Chambers (1984) considered the function (9.6), the division of \mathbf{x}_i into two components, and studied the behavior of the estimator from a model perspective. The procedure associated with (9.5), (9.6) and (9.7) was used by Isaki, Tsay and Fuller (2000). In that application, the vector $\bar{\mathbf{x}}_{0,N}$ contained marginal totals of a multiway table and $\bar{\mathbf{x}}_{2,N}$ contained totals for interior cells. Rao and Singh (1997) studied a closely related estimator in which tolerances are given for the difference between the final estimates for elements of $\bar{\mathbf{x}}_{2,N}$ and the corresponding elements of $\bar{\mathbf{x}}_{2,N}$.

Park (2002) extended Husain's optimality results to a more general Ψ . The \mathbf{x}_2 vector can be transformed so that $\tilde{\mathbf{V}}\{\bar{\mathbf{x}}_{2,\pi}\}$ for the transformed vector is a diagonal matrix and so that $\tilde{\mathbf{X}}_2' \Phi_{ee}^{-1} \tilde{\mathbf{X}}_2$ is a diagonal matrix, where $\tilde{\mathbf{X}}_2$ is the part of \mathbf{X}_2 that is orthogonal to \mathbf{X}_0 in the metric Φ_{ee} . That is,

$$\tilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_0(\mathbf{X}_0' \Phi_{ee}^{-1} \mathbf{X}_0)^{-1} \mathbf{X}_0' \Phi_{ee}^{-1} \mathbf{X}_2.$$

Then the diagonal Ψ that minimizes the approximate variance has elements

$$\psi_{ii} = (m_{ii} V_{\beta\beta ii})^{-1} \beta_i^2, \quad (9.12)$$

where m_{ii} is the i th element of the diagonal matrix $\tilde{\mathbf{X}}_2' \Phi_{ee}^{-1} \tilde{\mathbf{X}}_2$ and $V_{\beta\beta ii}$ is the variance of $\hat{\beta}_i$ in the transformed scale. To implement the procedure one must estimate the population parameters or choose realistic values for a general purpose Ψ . If one postulates a super-population random model for $\boldsymbol{\beta}$, then the β_i^2 of (9.12) is replaced with $E\{\beta_i^2\}$, where the expectation is the model expectation.

10. COMMENTS

Regression estimation is a flexible and powerful tool for the incorporation of auxiliary information into the estimation process. Closely related procedures, such as raking ratio, have large sample properties equivalent to those of regression estimators. The linearity of such estimators is of paramount importance because it permits the construction of a general purpose data set that provides very good estimators for a wide range of parameters.

Given a concentrated interest in a single y -variable, efficiency gains may be possible by postulating a particular set of auxiliary variables and a particular error covariance matrix. Because of the simple nature of the design consistency requirement, it is easy to test such models for design consistency.

ACKNOWLEDGEMENTS

This research was partially supported by Cooperative Agreement 43-3AEU-0-80064 between Iowa State University, the U. S. National Agricultural Statistics Service and the U. S. Bureau of the Census. I am deeply indebted to Mingue Park for assistance in literature review, for comments on and repair of theoretical results, and for use of material from his thesis. I thank Michael Hidioglou, J.N.K. Rao, Harold Mantel, and Jean Opsomer for useful comments on drafts of the manuscript. I thank the Associate Editor for numerous comments that improved the presentation.

APPENDIX

This appendix contains theorems supporting the limiting properties of the regression estimators discussed in section 4.

Theorem A.1. Let $\{U_N, F_N, A_N, n_N: N = k + 3, k + 4, \dots\}$ be a sequence of finite populations and samples, where F_N is a sample from an infinite population with eighth moments, A_N is the sample of size n_N selected from the N th population. Let $\hat{\boldsymbol{\beta}}$ be defined by (4.4) of the text, and let

$$\hat{\mathbf{Q}}_{zz} = \mathbf{n}^{-1} \mathbf{Z}' \Phi^{-1} \mathbf{Z},$$

where Φ is a positive definite symmetric $n \times n$ matrix that may be a function of \mathbf{X} but not of \mathbf{y} , \mathbf{Z} is defined following (4.2), and we omit the subscript N on sample quantities. Assume $\hat{\mathbf{Q}}_{zz}$ is positive definite with probability one. If Φ is random, assume the rows of $\Phi^{-1} \mathbf{Z}$ have bounded fourth moments. Assume the selection probabilities satisfy

$$0 < K_1 < N n^{-1} \pi_i < K_2,$$

where π_i are the selection probabilities. Assume the sample design is such that for any \mathbf{z} with bounded fourth moments

$$[(\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N)', (\hat{\mathbf{Q}}_{zz} - \mathbf{Q}_{zzN})] | \mathbf{F}_N = O_p(n^{-1/2}), \quad (\text{A.1})$$

where

$$\bar{\mathbf{z}}_{\text{HT}} = (\bar{y}_{\text{HT}}, \bar{\mathbf{x}}_{\text{HT}})' = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i, \quad (\text{A.2})$$

$\mathbf{Q}_{zzN} = E\{\hat{\mathbf{Q}}_{zz} | \mathbf{F}_N\}$, $\bar{\mathbf{z}}_N$ is the finite population mean of \mathbf{z} , \mathbf{Q}_{zzN} is a positive definite matrix for the N th population, and the limit of \mathbf{Q}_{zzN} is positive definite. Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N | \mathbf{F}_N = \mathbf{Q}_{xxN}^{-1} \bar{\mathbf{b}}'_{\text{HT}} + O_p(n^{-1}), \quad (\text{A.3})$$

where $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1} \mathbf{Q}_{xyN}$, $\bar{\mathbf{b}}_{\text{HT}} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i$, $\mathbf{b}_i' = n^{-1} N \pi_i \zeta_i' e_i$,

$$\mathbf{Q}_{zzN} = \begin{pmatrix} \mathbf{Q}_{yyN} & \mathbf{Q}_{yxN} \\ \mathbf{Q}_{xyN} & \mathbf{Q}_{xxN} \end{pmatrix}, \quad (\text{A.4})$$

$e_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}_N$, and ζ_i' is column i of $\mathbf{X}' \Phi^{-1}$. Assume the design is such that

$$\mathbf{V}_{\bar{\mathbf{z}}}^{-1/2} \{\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N | \mathbf{F}_N\} \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad (\text{A.5})$$

as $n_N \rightarrow \infty$ for any \mathbf{z} with finite fourth moments, where $\mathbf{V}_{\bar{\mathbf{z}}}$ is the covariance matrix of $\bar{\mathbf{z}}_{\text{HT}} - \bar{\mathbf{z}}_N$. Assume that $\mathbf{V}_{\bar{\mathbf{z}}}$ is $O(n^{-1})$ and that the design admits an estimator $\hat{\mathbf{V}}_{\bar{\mathbf{z}}}$ such that

$$n(\hat{\mathbf{V}}_{\bar{\mathbf{z}}} - \mathbf{V}_{\bar{\mathbf{z}}}) | \mathbf{F}_N = o_p(1) \quad (\text{A.6})$$

for any \mathbf{z} with bounded fourth moments. Then

$$[\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\}]^{-1/2} [\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N] | \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad (\text{A.7})$$

where

$$\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\} = \hat{\mathbf{Q}}_{xx}^{-1} \hat{\mathbf{V}}_{\bar{\mathbf{b}}} \hat{\mathbf{Q}}_{xx}^{-1}, \quad (\text{A.8})$$

$\hat{\mathbf{V}}_{\bar{\mathbf{b}}} = \hat{\mathbf{V}}\{\bar{\mathbf{b}}_{\text{HT}}\}$ is the estimated design variance of $\bar{\mathbf{b}}'_{\text{HT}}$ calculated with $\bar{\mathbf{b}}'_i = n^{-1} N \pi_i \zeta_i' \hat{e}_i$ and $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$.

Proof. The error in $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= (\mathbf{X}' \Phi^{-1} \mathbf{X})^{-1} [\mathbf{X}' \Phi^{-1} \mathbf{y} - \mathbf{X}' \Phi^{-1} \mathbf{X} \boldsymbol{\beta}_N] \\ &= \hat{\mathbf{Q}}_{xx}^{-1} (n^{-1} \mathbf{X}' \Phi^{-1} \mathbf{e}). \end{aligned}$$

Now $\hat{\boldsymbol{\beta}}$ is a generalized least squares estimator. Therefore

$$\hat{\mathbf{e}}' \Phi^{-1} \mathbf{X} = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})' \Phi^{-1} \mathbf{X} = \mathbf{0}$$

and $\mathbf{Q}_{xyN} - \boldsymbol{\beta}_N' \mathbf{Q}_{xxN} = \mathbf{Q}_{exN} = \mathbf{0}$. By assumption (A.1)

$$\hat{\mathbf{Q}}_{ex}' = n^{-1} \mathbf{X}' \Phi^{-1} \mathbf{e} = O_p(n^{-1/2}).$$

Thus

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N &= \mathbf{Q}_{xxN}^{-1} \left(n^{-1} \sum_{i \in A} \zeta_i' e_i \right) + O_p(n^{-1}) \\ &= \mathbf{Q}_{xxN}^{-1} \left(N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i' \right) + O_p(n^{-1}). \end{aligned}$$

The \mathbf{b}_i have bounded fourth moments by the assumptions. Thus, by assumption (A.5)

$$\mathbf{V}_{\bar{\mathbf{b}}}^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}),$$

where

$$\mathbf{V}_{\bar{\mathbf{b}}} = \mathbf{Q}_{xxN}^{-1} \mathbf{V}_{\bar{\mathbf{b}}} \mathbf{Q}_{xxN}^{-1}$$

and $\mathbf{V}_{\bar{\mathbf{b}}} = V\{\bar{\mathbf{b}}_{\text{HT}}\}$. Now

$$\begin{aligned} n^{-1} \mathbf{X}' \Phi^{-1} \hat{\mathbf{e}} &= n^{-1} \mathbf{X}' \Phi^{-1} \mathbf{e} + n^{-1} \mathbf{X}' \Phi^{-1} \mathbf{X} (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}) \\ &=: N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i' + N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{h}_i', \end{aligned}$$

where

$$\mathbf{h}_i' = n^{-1} N \pi_i \zeta_i' \mathbf{x}_i \delta_{\boldsymbol{\beta}}$$

and $\delta_{\boldsymbol{\beta}} = \boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}$. For any fixed δ , by (A.6), the estimated variance of $N^{-1} \sum_{i \in A} \pi_i^{-1} (\mathbf{b}_i' + \mathbf{h}_i')$ is consistent for the variance of the estimator of the mean of $\mathbf{b} + \mathbf{h}$. By assumption, the elements of $\zeta_i' \mathbf{x}_i$ have fourth moments. For a fixed δ the variance of $\bar{\mathbf{h}}_{\text{HT}}$ is $O(n^{-1})$. For $\delta = \delta_{\boldsymbol{\beta}}$,

$$\hat{\mathbf{V}}\{\bar{\mathbf{h}}_{\text{HT}}'\} = o_p(n^{-1}),$$

and

$$\hat{\mathbf{V}}\{\bar{\mathbf{b}}_{\text{HT}}'\} = V\{\bar{\mathbf{b}}_{\text{HT}}'\} + o_p(n^{-1})$$

because $\delta_{\boldsymbol{\beta}} = O_p(n^{-1/2})$. Result (A.7) then follows from the asymptotic normality of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N$.

Theorem A.2. Let $\mathbf{y}' = (y_1, y_2, \dots, y_n)$ and $\mathbf{X}' = (\mathbf{x}_1', \mathbf{x}_2', \dots, \mathbf{x}_n')$. Let Φ be a nonsingular symmetric $n \times n$ matrix and let Φ_N be a nonsingular symmetric $N \times N$ matrix. Let

$$\bar{y}_N, \bar{\mathbf{x}}_N, n^{-1} (\mathbf{X}' \Phi^{-1} \mathbf{X}) \text{ and } n^{-1} \mathbf{X}' \Phi^{-1} \mathbf{y}$$

be design consistent estimators for finite population characteristics $\bar{y}_N, \bar{\mathbf{x}}_N, \mathbf{Q}_{xxN}$ and \mathbf{Q}_{xyN} , respectively, where

$$[\mathbf{Q}_{xxN}, \mathbf{Q}_{xyN}] = \left[N^{-1} \mathbf{X}_N' \mathbf{\Phi}_N^{-1} \mathbf{X}_N, N^{-1} \mathbf{X}_N' \mathbf{\Phi}_N^{-1} \mathbf{y}_N \right]. \quad (\text{A.9})$$

Let $\boldsymbol{\beta}_N = \mathbf{Q}_{xxN}^{-1} \mathbf{Q}_{xyN}$. Let there be a sequence of column vectors $\{\boldsymbol{\gamma}_N\}$ such that

$$\mathbf{X} \boldsymbol{\gamma}_N = \mathbf{\Phi} \mathbf{D}_\pi^{-1} \mathbf{J}_n \quad (\text{A.10})$$

for all possible samples, where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \dots, \pi_n)$ and \mathbf{J}_n is an n -dimensional column vector of ones. Then, the regression estimator $\bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}$ with

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{\Phi}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{\Phi}^{-1} \mathbf{y}, \quad (\text{A.11})$$

is a design consistent estimator of $\bar{\mathbf{y}}_N$.

Proof. If $\hat{\boldsymbol{\beta}}$ is defined by (A.11), then by the properties of generalized least squares estimators,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{\Phi}^{-1} \mathbf{X} = \mathbf{0}.$$

If (A.10) holds, then

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{D}_\pi^{-1} \mathbf{J} = \left(\sum_{i \in A} \pi_i^{-1} \right) (\bar{y}_\pi - \bar{\mathbf{x}}_\pi \hat{\boldsymbol{\beta}}) = 0.$$

It follows that \bar{y}_{reg} is design consistent because

$$\begin{aligned} 0 &= p \lim_{N \rightarrow \infty} \left\{ (\bar{y}_\pi - \bar{\mathbf{x}}_\pi \hat{\boldsymbol{\beta}}_N) | \mathbf{F}_N \right\} \\ &= p \lim_{N \rightarrow \infty} \left\{ (\bar{y}_\pi - \bar{\mathbf{x}}_\pi \boldsymbol{\beta}_N) | \mathbf{F}_N \right\} \\ &= p \lim_{N \rightarrow \infty} \left\{ (\bar{y}_N - \bar{\mathbf{x}}_N \boldsymbol{\beta}_N) | \mathbf{F}_N \right\}. \end{aligned}$$

Theorem A.3. Let a sequence of populations and samples be as defined in Theorem A.1. Let \mathbf{z}_i be a vector of the form $\mathbf{z}_i = (y_i, 1, \mathbf{x}_{1,i})$ and let $\mathbf{z}_{1,i} = (y_i, \mathbf{x}_{1,i})$. Assume $\bar{\mathbf{z}}_{1,\pi}$ is a design consistent estimator of the population mean $\bar{\mathbf{z}}_{1,N}$ with nonsingular covariance matrix

$$V\{\bar{\mathbf{z}}_{1,\pi} | \mathbf{F}_N\} = O(n^{-1}) \quad (\text{A.12})$$

and

$$n^{1/2}(\bar{\mathbf{z}}_{1,\pi} - \bar{\mathbf{z}}_{1,N}) | \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Sigma}_{\bar{\mathbf{z}}}), \quad (\text{A.13})$$

where $\boldsymbol{\Sigma}_{\bar{\mathbf{z}}}$ is the limit of $n V\{\bar{\mathbf{z}}_{1,\pi} | \mathbf{F}_N\}$. Assume there is an estimator of the variance of $\bar{\mathbf{z}}_{1,\pi}$, denoted by $\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\}$, such that

$$p \lim_{N \rightarrow \infty} n^{1+\delta} \left(\hat{V}\{\bar{\mathbf{z}}_{1,\pi}\} - V\{\bar{\mathbf{z}}_{1,\pi} | \mathbf{F}_N\} \right) = \mathbf{0} \quad (\text{A.14})$$

for some $\delta > 0$. Let $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$ be the vector that minimizes

$$\hat{V}\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1,d}\} \quad (\text{A.15})$$

and let $\boldsymbol{\beta}_{1,\text{dopt}}$ be the vector that minimizes $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}_{1,d}\}$. Let $\bar{y}_{d,\text{reg}}$ be defined by (4.29). Then $\bar{y}_{d,\text{reg}}$ has the minimum limit variance for design consistent estimators of the form $\bar{y}_\pi + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \boldsymbol{\beta}_{1,d}$. Also

$$\left[\hat{V}\{\bar{e}_\pi\} \right]^{-1/2} (\bar{y}_{d,\text{reg}} - \bar{y}_N) \xrightarrow{L} N(0, 1), \quad (\text{A.16})$$

where $\hat{V}\{\bar{e}_\pi\}$ is the estimator of (A.14) constructed with $\hat{e}_i = y_i - \bar{y}_\pi - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_{1,\text{dopt}}$.

Proof. The estimator

$$\hat{\boldsymbol{\beta}}_{1,\text{dopt}} = \left[\hat{V}\{\bar{\mathbf{x}}_{1,\pi}\} \right]^{-1} \hat{C}\{\bar{\mathbf{x}}_{1,\pi}, \bar{y}_\pi\}$$

minimizes the estimated variance of (A.15), and, by assumption (A.14), the estimated variance is consistent for the true variance. Hence, $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$ is design consistent for $\boldsymbol{\beta}_{1,\text{dopt}}$ and $\hat{\boldsymbol{\beta}}_{1,\text{dopt}}$ minimizes $V\{\bar{y}_\pi - \bar{\mathbf{x}}_{1,\pi} \boldsymbol{\beta}\}$. Therefore, no estimator of the form (4.29) has a limit distribution with smaller variance.

Now

$$\begin{aligned} \bar{y}_{d,\text{reg}} - \bar{y}_N &= \bar{y}_\pi - \bar{y}_N - (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,\pi}) \hat{\boldsymbol{\beta}}_{1,\text{dopt}} \\ &= \bar{e}_\pi + o_p(n^{-1/2}), \end{aligned}$$

where $e_i = y_i - \bar{y}_N - (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N}) \boldsymbol{\beta}_{1,\text{dopt}}$. Therefore the variance of the limiting distribution of $n^{1/2}(\bar{y}_{d,\text{reg}} - \bar{y}_N)$ is the variance of $n^{1/2}(\bar{e}_\pi - \bar{e}_N)$. By assumption (A.14), the estimator $\hat{V}\{\bar{\mathbf{z}}_\pi \boldsymbol{\gamma}\}$ is a consistent variance estimator of $V\{\bar{\mathbf{z}}_\pi \boldsymbol{\gamma}\}$ for any fixed $\boldsymbol{\gamma}$. Because $\hat{\boldsymbol{\beta}}_{1,\text{dopt}} - \boldsymbol{\beta}_{1,\text{dopt}} = o_p(1)$, the estimated variance based on \hat{e}_i converges to the estimated variance based on e_i and (A.16) holds.

REFERENCES

- ANDERSON, C., and NORDBERG, L. (1998). A user's guide to CLAN97. Statistics Sweden, Orebro, Sweden.
- BANKIER, M.D., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working Paper-Methodology Branch, Census Operations Section, Social Survey Methods Division. Statistics Canada, Ottawa.
- BANKIER, M.D., HOULE, A.M. and LUC, M. (1997). Calibration estimation in the 1991 and 1996 Canadian census. Statistics Canada (draft), 8 pages.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.
- BREWER, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

- BREWER, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- BREWER, K.R.W., HANIF, M. and TAM, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.
- BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.
- CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1979). Prediction theory for finite populations when model-based and design-based principles are combined. *Scandinavian Journal of Statistics*, 6, 97-106.
- CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, (Eds. W.G. Madow, I. Olkin, and D. Rubin). New York: Academic Press, 3, 143-160.
- CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, Inc.
- COOK, R.D., and WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEMING, W.E., and STEPHAN, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 45-49.
- DEVILLE, J., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J., SÄRNDAL, C.-E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DORFMAN, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.
- DORFMAN, A.H., and HALL, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1475.
- DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.
- DU MOUCHEL, W. H., and DUNCAN, G. J. (1983). Using survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- ELTINGE, J.L., and YANSANEH, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U. S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.
- ESTEVAO, V., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.
- FOLSOM, R.E., and WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 428-433.
- FOLSOM, R.E., and SINGH, A.C. (2000). The generalized exponential model for a unified approach to sampling weight calibration for outlier weight treatment, nonresponse adjustment and post-stratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 598-603.
- FRANKEL, M.R. (1971). Inference from survey samples: An empirical investigation. Institute for Social Research, University of Michigan, Ann Arbor.
- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- FULLER, W.A. (1973). Regression for sample surveys. Paper presented at meeting of International Statistical Institute. August, 1973, Vienna, Austria.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā Series C*, 37, 117-132.
- FULLER, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.
- FULLER, W.A., and AN, A.B. (1998). Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.

- FULLER, W.A., and BATTESE, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.
- FULLER, W.A., and ISAKI, C.T. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling*, (D. Krewski, J. N. K. Rao and R. Platek, Eds.), New York: Academic Press, 199-226.
- FULLER, W.A., LOUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey, *Survey Methodology*, 20 75-85.
- FULLER, W.A., and RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology*, 27, 45-52.
- GAMBINO, J., KENNEDY, B. and SINGH, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.
- GEROW, K., and MCCULLOCH, C.E. (2000). Simultaneously model unbiased, design-unbiased estimation. *Biometrics* 56, 873-878.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.
- GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.
- GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369-375.
- GRAYBILL, F.A. (1976). *Theory and application of the linear model*. Wadsworth, Belmont, CA.
- HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pestovani Matematiky*, 84, 387-423.
- HANURAV, T.V. (1966). Some aspects of unified sampling theory. *Sankhyā, Series A*, 28, 175-204.
- HENDERSON, C.R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, 141-163. National Academy Sciences, National Research Council Publication 982, Washington, DC.
- HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Ph.D. thesis, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1978). *Super Carp*, (sixth edition, 1980) Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., SÄRNDAL, C.-E. and BINDER, D. A. (1995). Weighting and estimation in business surveys. *Business Survey Methods*, (Eds. Cox, Binder, Chinnappa, Colledge and Kot) New York: John Wiley & Sons, Inc., 477-502.
- HOLT, D., and SMITH, T.M. F. (1979). Post Stratification. *Journal of the Royal Statistical Society, Serie. A*, 142, 33-46.
- HORN, S.D., HORN, R.A. and DUNCAN, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the social statistics section*, American Statistical Association, 300-305.
- HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.
- ISAKI, C.T. (1970). Survey designs utilizing prior information. Unpublished Ph.D. thesis. Iowa State University.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- ISAKI, C.T., TSAY, J.H. and FULLER, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*, 26, 31-42.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agriculture Experiment Station Research Bulletin*. 304
- KALTON, G., and MALIGALIG, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference*, U. S. Bureau of the Census, 409-428.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KONIJN, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association*, 57, 590-606.
- KOTT, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.
- KUO, L. (1988). Classical and prediction approaches to estimating distribution function from survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 280-285.
- LAZZERONI, L.C., and LITTLE, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

- MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- MONTANARI, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.
- MONTANARI, G.E. (1999). A study on the conditional properties of finite population mean estimators. *Metron*, 57, 21-35.
- MUKHOPADHYAY, P. (1993). Estimation of a finite population total under regression models: A review. *Sankhyā*, 55, 141-155.
- NIEUWENBROEK, N., RENSSSEN, R. and HOFMAN, L. (2000). Towards a generalized weighting system. In *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia.
- PARK, M. (2002). Regression estimation of the mean in Survey Sampling. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.
- PFEFFERMANN, D. (1984). Note on large sample properties of balanced samples. *Journal of the Royal Statistical Society, Series B*, 46, 38-41.
- RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- RAO, J.N.K. (2002). *Small Area Estimation Theory and Methods*, New York: John Wiley & Sons, Inc.
- RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.
- RAO, J.N.K., and SINGH, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*, American Statistical Association, 57-64.
- ROBINSON, G.K. (1991). The BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-32.
- ROBINSON, P.M., and SÄRNDAL, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā, Series B*, 45, 240-248.
- ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika*, 70, 41-55.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALL, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Association*, 71, 657-664.
- ROYALL, R.M. (1986). The prediction approach to robust variance estimation in two stage cluster sampling. *Journal of the American Statistical Association*, 81, 119-123.
- ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). The finite population linear regression estimator and estimators of its variance, an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRNDAL, C.-E. (1980). On π -weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistics Association*, 91, 1289-1300.
- SÄRNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SÄRNDAL, C.-E., and WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- SCOTT, A., and SMITH, T.M.F. (1974). Linear superpopulation models in survey and sampling. *Sankhyā, C*, 36, 143-146.
- SCOTT, A., and WU, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.
- SILVA, P.L.D.N., and SKINNER, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.
- SINGH, A.C., and FOLSOM, R.E. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 610-615.
- SINGH, A.C., KENNEDY, B. and WU, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating design. *Survey Methodology*, 27, 33-44.
- SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.
- TALLIS, G.M. (1978). Note on robust estimation infinite populations. *Sankhyā, C*, 40, 136-138.
- TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.
- THÉBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.
- THÉBERGE, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26, 99-107.
- TILLE, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.
- TREMBLAY, V. (1986). Practical Criteria for Definition of Weighting Classes. *Survey Methodology*, 12, 85-97.

- WATSON, D. J. (1937). The estimation of leaf area in field crops. *Journal of Agricultural Science*, 27, 474-483.
- WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.
- WU, C., and SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- YATES, F. (1949). *Sampling Methods for Census and Surveys*. London: Griffin.
- YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.
- ZYSKIND, G. (1976). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38, 1092-1109.

Leslie Kish's Impact on Survey Statistics

GRAHAM KALTON¹

ABSTRACT

Leslie Kish, one of the pioneers of survey sampling, died on October 7, 2000, at the age of 90. This paper reviews his impact on survey statistics, mainly in terms of his research but also in terms of his promotion of sound probability sampling methods around the world. Kish's research was broad-ranging, covering sampling methods, variance estimation and design effects, nonsampling errors, small area estimation, survey designs across time and space, and observational studies. He promoted probability sampling designs through consultancies in many countries, his writings, and in particular through the highly effective intensive summer Sampling Program for Foreign Statisticians that he established at the Survey Research Center of the University of Michigan.

KEY WORDS: Sample design; Variance estimation; Nonsampling errors; Rolling samples.

1. INTRODUCTION

Leslie Kish, one of the pioneers of survey sampling, died on October 7, 2000, at the age of 90. During his long and productive career, he had a major impact on the field, achieved both through his impressive research contributions and through his extremely successful promotion of the use of scientific probability sampling methods throughout the world, and especially in developing countries. His wide-ranging research always focused on issues of practical importance, and his innovations facilitated the use of effective probability sampling in diverse areas. He promoted the practice of probability sampling through his expository writings (particularly for sociologists and demographers), through his numerous consultancies and advisory services, and through his training of survey statisticians, particularly those from developing countries.

This paper reviews Kish's impact on survey statistics, primarily with respect to his contributions to the advancement of survey sampling and survey research more generally. It is useful to start with a brief account of his career in order to place these contributions in a temporal context. The interview of Kish in 1994 by Frankel and King (1996) is recommended for those interested in more details of Kish's fascinating life. Some of the material in this paper is drawn from that interview.

Kish was born in 1910 in Poprad, which was then part of the Austro-Hungarian Empire and is now in Slovakia. In 1926, he emigrated to the United States with his family. When his father died the following year, he became a laboratory assistant at the Rockefeller Institute for Medical Research, while attending Bay Ridge Evening High School. He graduated from high school in 1930 and enrolled in the College of the City of New York night school, while continuing to work for 54 hours a week at the Rockefeller Institute. His interest in statistics arose out of his work at the Institute, and he studied on his own books by Fisher,

Yule, Wallace and Snedecor, Tippet, Pearl, and others. In 1937, he interrupted his education to join the International Brigade to fight for the Loyalist cause in the Spanish Civil War. He returned to the United States in 1939 and earned a B.S. in Mathematics, cum laude, in that year. He was then hired by the U.S. Census Bureau as a Section Head, and subsequently moved to be a Statistician at the United States Department of Agriculture (USDA) Division of Program Surveys. In 1942, he left the Division of Program Surveys for war service, returning there in 1945 after the war. In 1947, he moved with a group of USDA colleagues headed by Rensis Likert to set up the Survey Research Center at the University of Michigan. He remained at the Survey Research Center until his retirement in 1981, when he became a Professor Emeritus. He remained fully active professionally until his death in 2000.

2. RESEARCH

At the start of Kish's career, survey sampling was in its infancy. Much survey research was based on nonprobability samples. Methods for probability sampling were under development and many problems remained to be resolved. While at the USDA, Kish identified three important problems that he pursued at the Survey Research Center (SRC) in developing sampling methods there.

One of these problems was how to have an interviewer randomly select an individual within a sampled household. At the time, probability sampling methods for sampling households had been developed and were being applied in the Current Population Survey, but the CPS collected data on all members of sampled households, so that no selection of persons within households was needed. Kish invented a method for objective respondent selection and wrote it up in a memorandum. He was urged by his colleague Angus Campbell to submit the work for publication, and it resulted

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

in the famous paper that was his first published research (Kish 1949). The widely used method is now known as the Kish selection table.

The second problem that Kish identified was counting nonresponse. He had to argue for counting and reporting nonresponse with probability samples against the concerns of colleagues who felt that to do so would put the SRC at a competitive disadvantage, particularly with organizations using nonprobability methods. He won his case and SRC adopted his approach, which is now fully accepted as standard good practice.

The third problem was that of deep stratification. Standard stratification assumes independence of selections between strata, with the maximum number of strata possible being the number of selections. Particularly when the number of selections is small, as is often the case with the primary sampling units (PSUs) in a multistage design, it can be desirable to obtain greater balance in the sample than standard stratification permits. With Roe Goodman, Kish developed the technique of controlled selection that provides that greater balance by dropping the requirement of independence of selections between strata, while still retaining probability sampling (Goodman and Kish 1950). Kish, who was always concerned to coin good names, preferred to call the technique 'multiple stratification', and he uses that term in his sampling text (Kish 1965a).

Kish's subsequent research in survey statistics was wide-ranging, covering many aspects of survey sampling, nonsampling errors, small area estimation, survey designs across time and space, and observational studies. His many contributions have had a major impact on the development of the practice of survey sampling and of survey research more generally. The following paragraphs outline some of his contributions organized by topic.

Variance estimation. Before the 1970s, the analysis of survey data was severely limited by the analytic tools available, then mostly punch card equipment, such as counter-sorters and tabulators, and hand calculators. Thus, for example, weights – and particularly non-integer weights – were difficult to handle. For this reason Kish examined the use of uniform weights with the Kish selection table, even though unbiased estimation calls for weights proportional to the number of eligible household members.

As a result of the computational difficulties, prior to the 1970s sampling errors were rarely computed in a manner that reflected the complex sample designs typically employed in survey research. A widespread practice was to compute variances as if a simple random sample (SRS) had been drawn. Kish sought to promote the use of appropriate variance estimation methods by social researchers, which he did by illustrating the sizable underestimation that often arises when SRS formulas are applied to clustered samples (Kish 1957). Initially he developed and applied simple computational procedures, emphasizing the simplicity that can be obtained with a paired selection design in which two PSUs are sampled in each stratum (Kish and Hess 1959a;

Kish 1968). He coined the term "design effect" for the ratio of the variance of a survey estimate for a given design to the variance of the same estimate obtained from a simple random sample of the same size. He made much use of this concept in his famous *Survey Sampling* book (Kish 1965a), which provides an encyclopedic treatment of practical survey sampling and is still widely read as a Wiley classic. He retained his interest in design effects throughout his career as an important tool in the design and analysis of survey samples (see, for example, Kish 1982, 1995a; Kish, Frankel, Verma and Kaciroti 1995; Kish, Groves and Krotki 1976). An important term in the design effect for a clustered sample is the intra-class correlation, which is featured in Kish's Ph.D. dissertation (Kish 1952) and in a number of his other papers (e.g., Kish 1954, 1961a).

With the development of computers, Kish was quick to see their importance for variance estimation, and with SRC colleagues he developed an early *Sampling Error Program Package* (Kish, Frankel and Van Eck 1972). With his doctoral student Martin Frankel, he also extended the range of statistics for which sampling errors from complex sample designs could be computed (Kish and Frankel 1970, 1974). This highly influential research developed, applied, and evaluated balanced repeated replication (BRR) and jackknife repeated replication (JRR) methods of variance estimation. It also provided a definition of the population parameters estimated by analytical survey statistics in the finite population context.

Multipurpose surveys. The survey sampling literature deals mostly with an efficient sample design for estimating a single population parameter. Kish recognized the limitation of this approach since virtually all surveys are multipurpose in nature. He wrote several important papers dealing with multipurpose surveys, producing effective compromise designs that provide estimates not only for the population as a whole but also for various domains (Kish 1961b, 1969, 1976; Anderson, Kish and Cornell 1976; Kish and Anderson 1978; Kish 1980; Kish 1988). In recent years, he extended his interests to multipopulation surveys (e.g., Kish 1999, 2002).

Small area estimation. In considering the production of estimates for domains, Kish (1980, 1987a, 1987b) classified domains into major, minor, and mini domains and rare items. Estimates for major domains can be produced from a survey using standard sample-based estimators, particularly if the sample is designed to give sufficient domain sample sizes for this purpose. The sample sizes of most surveys preclude the production of estimates of adequate precision for minor or mini domains that comprise less than, say, one-tenth of the population. Yet, as Kish recognized early on, the demand for up-to-date estimates for small domains, particularly small geographical areas, would expand. This recognition led to his research in two related areas.

When a survey's sample size is too small to produce small area sample-based estimates of adequate precision,

reliance may be placed on statistical models to produce indirect estimates. Much research on small area estimation techniques using this model-dependent approach has been conducted in recent years. In the 1970's, Kish contributed to the development of the field through his direction of three doctoral dissertations at the University of Michigan (Erickson 1973; Kalsbeek 1973; Purcell and Kish 1979, 1980).

Direct, or sample-based, estimates for small domains are sometimes possible. One obvious source of estimates for domains of any size is a population census, and indeed censuses are a major source of small domain estimates. However, data from a decennial census become out-of-date as the decade progresses. To address this problem, Kish proposed replacing the census by a rotating or rolling sample so that, by spreading the data collection over time, more up-to-date estimates can be produced. He first proposed such a procedure in 1979 (Kish 1979a,b), and wrote many papers on this topic after that (Kish 1981, 1983, 1986, 1990, 1997, 1998, 2002; Kish and Verma 1986), including the issue of how to cumulate sample data over time (Kish 1999). In another paper in this volume, Charles Alexander (2002) provides a detailed review of Kish's work on this topic and its influence on the large-scale continuous survey, the American Community Survey, that the U.S. Census Bureau plans to introduce to replace the long form in the 2010 Census.

Special sample design problems. During the course of his work, Kish encountered a number of specialized sampling problems that often occur and he offered some efficient solutions. The areas to which he contributed include the following:

- *Sampling rare and elusive populations.* One of the most challenging design tasks faced by sampling statisticians is constructing an efficient sample design for a rare or elusive population (such as persons with a rare illness or the homeless). Kish (1965b, 1991) provides insightful reviews of methods for tackling this type of problem.
- *Maximizing overlap.* When a population is sampled repeatedly over time, the issue arises of how to control the sample overlap between one round and the next. A particular example occurs when a master sample of PSUs is used and needs to be updated when new census data become available. Frequently it is desirable to maximize the overlap in the sample of PSUs, while updating measures of size and changing the stratification to reflect current data. Kish and Scott (1971) provide a relatively simple and effective method of satisfying these requirements.
- *Sampling organizations of unequal size.* Some surveys are designed to produce estimates for units at different levels, for instance, for hospitals and

for patients. When hospitals vary considerably in their numbers of patients, a design conflict arises between the production of efficient hospital- and patient-level estimates. Kish (1965c) examines this problem and clarifies the issues involved.

Nonsampling errors. Kish clearly recognized the harmful effects that nonsampling errors can have on the quality of survey estimates. Early in his career he collaborated with Jack Lansing to investigate the response errors in respondents' reports of the values of their homes by comparing these reports with estimates made by professional appraisers (Kish and Lansing 1954). In his studies of interviewer variance, he took advantage of the theory on cluster sampling, measuring interviewer variance with the intra-class correlation coefficient, and determining the optimum number of interviews per interviewer based on a simple cluster sample cost model (Kish 1962). With Irene Hess, he conducted a study of noncoverage in area samples of dwelling units. The study was stimulated by a 10 percent noncoverage rate in SRC surveys at that time, and led to improvements that reduced this rate to about 3 percent (Kish and Hess 1958). Also with Irene Hess, he introduced an imaginative replacement procedure for noncontacts in one survey by substituting noncontacts from a previous, similar, survey (Kish and Hess 1959b). For stochastic imputation schemes, Kish was an early proponent of replicating the imputations to reduce imputation variance, in what he termed a repeated replication imputation procedure (RRIP) and what is now known as fractional imputation (Kalton and Kish 1984).

Observational studies. Early in his career, Kish (1959) wrote a widely cited paper on the design of studies to investigate causal relationships, particularly nonrandomized studies. In his writing about this topic he made use of his survey sampling expertise as, for instance, in the relationship between stratification and matching (Anderson, Kish and Cornell 1980). His work developed into his book *Statistical Design for Research* (Kish 1987a) in which he compared surveys, experiments, and observational studies for investigating causal effects in terms of the three R's: realism, randomization and representativeness (see also Kish 1975). He also made clear the importance of assessing both bias and variance in assessing the ability of different study designs to measure causal effects, rather than concentrating on bias as had been common in the literature on this topic.

3. OTHER CONTRIBUTIONS

Kish's seminal and wide-ranging contributions to the methodology of survey statistics are of great importance. Yet of possibly even greater importance are his contributions to the promotion of the use of sound probability sampling methods around the world.

Kish's writings, of course, contributed to the current widespread use of probability sampling methods by emphasizing good practical methods. His three books *Survey Sampling* (Kish 1965a), *Statistical Design for Research* (Kish 1987a), and *Sampling Methods for Agricultural Surveys* (Kish 1989) are all extremely valuable in this respect, as are his expository writings for social scientists.

Kish had a long-standing dedication to assisting developing and transition countries, and that can be seen in many of his activities. He was a sampling consultant to the World Fertility Survey from 1973 to 1983 and he consulted in many countries, he ran a training program for foreign statisticians, and he wrote specifically for statisticians in developing countries. *Sampling Methods for Agricultural Surveys* was, for instance, written for the FAO, particularly for use in developing countries. He contributed a *Questions/Answers* column for the *Survey Statistician*, the newsletter of the International Association of Survey Statisticians, from 1978 to 1994. In that column he provided sound advice on many practical sampling problems that frequently arise but that are not well addressed in the literature. The column was considered so useful that the IASS published the full set of questions and answers in a special volume (Kish 1995b).

Kish was rightly particularly proud of the intensive two-month summer Sampling Program for Foreign Statisticians that he established at the Survey Research Center in 1961. The SPFS has now trained more than 500 survey statisticians from 105 countries. It is significant that Kish chose "Developing samplers for developing countries" as the topic for his 1994 Morris Hansen Memorial Lecture (Kish 1996). To help maintain this important program, the Leslie Kish International Fellows Fund was established at the University of Michigan at a celebration of Kish's 90th birthday. Of all his accomplishments, the SPFS was the one that gave him greatest pleasure.

4. CONCLUDING REMARKS

Leslie Kish is a giant in the field of survey sampling. His contributions were enormous and recognized by many honors. These honors included, among others, President of the International Association of Survey Statisticians in 1983-85, President of the American Statistical Association in 1978 (see Kish 1978, for his Presidential address on "Chance, Statistics and Statisticians"), Honorary Fellow of the International Statistical Institute, Honorary Fellow of the Royal Statistical Society, Honorary Member of the Hungarian Academy of Sciences, Fellow of the American Association for the Advancement of Science, Fellow of the American Academy of Arts and Sciences, recipient of the American Statistical Association's Samuel L. Wilks Award in 1997, recipient of the Mindel Shep Award from the Population Association of America in 1998, recipient of the Methodology Award from the American Sociological

Association in 1989, and honorary degrees from the University of Bologna, the Athens University of Economics and Business, and the Eotvos Lorand University in Budapest.

Yet Kish remained down-to-earth, approachable by all. He had a great enthusiasm for many subjects including sport, art, literature, politics, philosophy, and science. He was always concerned with improving the conditions of the world's population. He was particularly interested in young people and one of his favorite sayings was "Keep young by being curious, and have young friends". Undoubtedly his endearing personality played an important part in his great success in promoting sound sampling methods around the world. Ivan Fellegi's excellent obituary in *Survey Methodology* was aptly titled "Leslie Kish – A Life of Giving" (Fellegi 2000). Kish gave so much personally to so many people and so much professionally to the development of survey statistics.

REFERENCES

- ALEXANDER, C. H. (2002). Still rolling: Leslie Kish's "rolling samples" and the American Community Survey. *Survey Methodology*, 28, 35-41.
- ANDERSON, D.W., KISH, L. and CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.
- ANDERSON, D.W., KISH, L. and CORNELL, R.G. (1980). On stratification, grouping, and matching. *Scandinavian Journal of Statistics*, 7, 61-66.
- ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.
- FELLEGI, I.P. (2000). Leslie Kish – A life of giving. *Survey Methodology*, 26, 119-120.
- FRANKEL, M., and KING, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 65-87.
- GOODMAN, R., and KISH, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- KALSBEEK, W.D. (1973). A Method for Obtaining Local Postcensal Estimates for Several Types of Variables. Ph. D. Thesis, University of Michigan.
- KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, 13(16), 1919-1939.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KISH, L. (1952). On the Differentiation of Ecological Units. Ph.D. Thesis, University of Michigan.
- KISH, L. (1954). Differentiation in metropolitan areas. *American Sociological Review*, 19, 388-398.
- KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 1954-1965.

- KISH, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24, 328-338.
- KISH, L. (1961a). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*, 4, 201-209.
- KISH, L. (1961b). Efficient allocation of a multi-purpose sample. *Econometrica*, 29, 363-385.
- KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- KISH, L. (1965a). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1965b). Selection techniques for rare traits. *Genetics, Epidemiology, and Chronic Diseases*, Public Health Service Publication, No. 1173.
- KISH, L. (1965c). Sampling organizations and groups of unequal sizes. *American Sociological Review*, 20, 564-572.
- KISH, L. (1968). Standard errors for indexes from complex samples. *Journal of the American Statistical Association*, 63, 512-529.
- KISH, L. (1969). Design and estimation for subclasses, comparisons, and analytical statistics. *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc.
- KISH, L. (1975). Representation, randomization and control. *Quantitative Sociology*, (Eds. H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon and V. Capecchi). New York: Academic Press.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, A*, 139, 80-95.
- KISH, L. (1978). Chance, statistics, and statisticians. *Journal of the American Statistical Association*, 73, 1-6.
- KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.
- KISH, L. (1979b). Rotating samples instead of censuses. *Asian and Pacific Census Forum* (East-West Center, Honolulu), 6, 1-13.
- KISH, L. (1980). Design and estimation for domains. *The Statistician*, 29, 209-222.
- KISH, L. (1981). Using Cumulated Rolling Samples. Washington: Library of Congress.
- KISH, L. (1982). Design effects. *Encyclopedia of Statistics*, New York: John Wiley & Sons, Inc.
- KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright). New York: Academic Press.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.
- KISH, L. (1987a). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.
- KISH, L. (1987b). Discussion. *Small Area Statistics*, (Ed. R. Platek). New York: John Wiley & Sons, Inc.
- KISH, L. (1988). Multipurpose sample design. *Survey Methodology*, 14, 19-32.
- KISH, L. (1989). *Sampling Methods for Agricultural Surveys*. Rome: FAO.
- KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-71 and 93-94.
- KISH, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339-347.
- KISH, L. (1995a). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.
- KISH, L. (1995b). *Questions/Answers from the Survey Statistician 1978-1994*. Libourne: International Association of Survey Statisticians.
- KISH, L. (1996). Developing samplers for developing countries. *International Statistical Review*, 64, 143-162.
- KISH, L. (1997). Periodic and rolling samples and censuses. *Statistics and Public Policy*, (Ed. B.D. Spencer). New York: Oxford University Press.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.
- KISH, L. (1999). Combining/cumulating population surveys. *Survey Methodology*, 25, 129-138.
- KISH, L. (2002). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, 102, 109-118.
- KISH, L., and ANDERSON, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 24-34.
- KISH, L., and FRANKEL, M. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, B*, 36, 1-37.
- KISH, L., FRANKEL, M.R. and VAN ECK, M. (1972). *SEPP: Sampling Error Programs Package*. Ann Arbor: Institute for Social Research.
- KISH, L., FRANKEL, M.R., VERMA, V. and KACIROTI, N. (1995). Design effects for correlated $(P_i - P_j)$. *Survey Methodology*, 21, 117-124.
- KISH, L., GROVES, R.M. and KROTKI, K. (1976). Sampling Errors for Fertility Surveys. Occasional Paper No. 17, World Fertility Survey.
- KISH, L., and HESS, I. (1958). On noncoverage of sample dwellings. *Journal of the American Statistical Association*, 53, 509-524.
- KISH, L., and HESS, I. (1959a). On variances of ratios and their differences in multi-stage samples. *Journal of the American Statistical Association*, 54, 416-446.
- KISH, L., and HESS, I. (1959b). A replacement procedure for reducing the bias of nonresponse. *The American Statistician*, 13, 17-19.
- KISH, L., and LANSING, J.B. (1954). Response error in estimating the value of homes. *Journal of the American Statistical Association*, 49, 520-538.
- KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- KISH, L., and VERMA, V. (1986). Complete censuses and samples. *Journal of Official Statistics*, 2, 381-96.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (small domains). *International Statistical Review*, 48, 3-18.

New Paradigms (Models) for Probability Sampling

LESLIE KISH¹

1. STATISTICS AS A NEW PARADIGM

In several sections I discuss new concepts in diverse aspects of sampling, but I feel uncertain whether to call them new paradigms or new models or just new methods. Because of my uncertainty and lack of self-confidence, I ask the readers to choose that term with which they are most comfortable. I prefer to remove the choice of that term from becoming an obstacle to our mutual understanding.

Sampling is a branch of and a tool for statistics, and the field of statistics was founded as a new paradigm in 1810 by Quetelet (Porter 1987; Stigler 1986). This was later than the arrival of some sciences: of astronomy, of chemistry, of physics. "At the end of the seventeenth century the philosophical studies of cause and chance... began to move close together... During the eighteenth and nineteenth centuries the realization grew continually stronger that aggregates of events may obey laws even when individuals do not." (Kendall 1968). The predictable, meaningful, and useful regularities in the behavior of population aggregates of unpredictable individuals were named "statistics" and were a great discovery.

Thus Quetelet and others computed national (and other) birth rates, death rates, suicide rates, homicide rates, insurance rates, *etc.* from individual events that are unpredictable. These statistics are basic to fields like demography and sociology. Furthermore, the ideas of statistics were taken later during the nineteenth century also into biology by Frances Galton and Karl Pearson, and into physics by Maxwell, and were developed greatly both in theory and applications.

Statistics and statisticians deal with the effects of chance events on empirical data. The mathematics of chance had been developed centuries earlier for gambling games and for errors of observation in astronomy. Also data have been compiled for commerce, banking, and government. But combining chance with real data needs a new theoretical view, a new paradigm. Thus statistical science and its various branches arrived late in history and in academia, and they are products of the maturity of human development (Kish 1985).

The populations of random individuals comprise the most basic concept of statistics. It provides the foundation for distribution theories, inferences, sampling theory, experimental design, *etc.* And the statistics paradigm differs fundamentally from the deterministic outlook of cause and effect, and of precise relations in the other sciences and mathematics.

2. THE PARADIGM OF SAMPLING

The Representative Method is the title of an important monograph, almost a century after the birth of statistics and over a century ago now, which is generally accepted as the birth of modern sampling (Kiaer 1895). That term has been used in several landmark papers since then (Jensen 1926; Neyman 1934; Kruskal and Mosteller 1979a, 1979b, 1979c, 1980). The last authors agree that the term "representative" has been used for so many specific methods and with so many meanings that it does not denote any single method. However, as Kiaer used it, and as it is still used generally, it refers to the aims of selecting a sample to represent a population specified in space, in time, and by other definitions, in order to make statistical inferences from the sample to that specified population. Thus a national representative sample demands careful operations for selecting the sample from all elements of the national population, not only from some arbitrary domain such as a "typical" city or province, or from some subset, either defined or undefined.

The scientifically accepted method for survey sampling is probability sampling, which assures known positive probabilities of selection for every element in the frame population. The frame provides the equivalent of listings of sampling units for each stage of selection. The sampling frame for the entire population is needed for mechanical operations of random selection. This is the basis for statistical inferences from the sample statistics to the corresponding population statistics (parameters) (Hansen, Hurwitz and Madow 1953a, 1953b). This insistence on inferences based on selections from frame populations is a different paradigm from the unspecified or model based approaches of most statistical analyses.

It took a half century from Kiaer's paper to the wide acceptance of survey sampling. In addition to neglect and passive resistance, there was a great deal of active opposition by national statistical offices which distrusted sampling methods to replace the complete counts of censuses. Some even preferred the "monograph method," which offered complete counts of a "typical" or "representative" province or district instead of randomly selected national sample (O'Muircheartaigh and Wong 1981). In addition to political opposition, there were also many opponents among academic disciplines, and among academic statisticians. The tide in favor of probability sampling turned with the report of the UN Statistical Commission led by Mahalanobis and Yates (United Nations

¹ Printing of this paper has been kindly authorized by Rhea Kish, 1050 Wall St. #9A, Ann Arbor, MI 48105, e-mail: rheakk@umich.edu.

Statistical Office 1950). Five influential textbooks between 1949 and 1954 started a flood of articles with both theory and wide applications.

The strength, the breadth, and the duration of resistance to the concepts and use of probability sampling of frame populations implies that this was a new paradigm that needed a new outlook both by the public and the professionals.

3. COMPLEX POPULATIONS

The need for strict probability selection from a population frame for inferences from the sample to a *finite* population is but one distinction of survey sampling. But even more important and difficult problems are caused by the complex distributions of the elements in all the populations. These complexities present a great contrast with the simple model of independence that is assumed, explicitly or implicitly, by almost all statistical theory, all mathematical statistics.

The assumption of independent or uncorrelated observations of variables or elements underlies mathematical statistics and distribution theory. We need not distinguish here between independently and identically distributed (IID) random variables and “exchangeability,” and “superpopulations.” The simplicity underlying each of those models is necessary for the complexities of the mathematical developments.

Simple models are needed and used for early stages and introductions in all the sciences: for example, perfect circular paths for the planets or $d = gt^2/2$ for freely dropping objects in frictionless situations. But those models fail to meet the complexities of the actual physical world. Similarly, independence of elements does not exist in any *population* whether human, animal, plant, physical, chemical, biological. The simple independent models may serve well enough for small *samples*; and the Poisson distribution of deaths by horsekicks in the Prussian Army in 43 years has often served as an example (precious because rare) (Fisher 1926).

There have also been attempts to construct theoretical populations of IID elements; perhaps the most famous was the classic “collective” of Von Mises (1931); but they do not correspond to actual populations. However, with great effort tables of random numbers have been constructed that have passed all tests. These have been widely used in modern designs of experiments and sample surveys. *Replication* and *randomization* are two of the most basic concepts of modern statistics following the concept of populations.

The simple concept of a population of independent elements does not describe adequately the complex distributions (in space, in time, in classes) of elements. Clustering and stratification are common names for ubiquitous complexities. Furthermore, it appears impossible

to form models that would better describe actual populations. The distributions are much too complex and they are also different for every survey variable. These complexities and differences have been investigated and presented now in thousands of computations of “design effects.”

Survey sampling needed a new paradigm to deal with the complexities of all kinds of populations for many survey variables and a growing list of survey statistics. This took the form of robust designs of selections and variance formulas that could use a multitude of sample designs, and gave rise to the new discipline of survey sampling. The computation of “design effects” demonstrated the existence, the magnitude, and the variability of effects due to the complexities of distributions not only for means but also for multivariate relations, such as regression coefficients. The long period of disagreements between survey samplers and econometricians testifies to the need for a new paradigm.

4. COMBINING POPULATION SAMPLES

Samples of national populations always represent subpopulations (domains) which differ in their survey characteristics; sometimes they differ slightly, but at other times greatly. These subclasses can be distinguished in the sample with more or less effort. First, samples of provinces are easily separated when their selections are made separately. Second, subclasses by age, sex, occupation, and education can also be distinguished, and sometimes used for poststratified estimates. Third, however, are those subclasses by social, psychological, and attitudinal characteristics, which may be difficult to distinguish; yet they may be most related to the survey variables. Thus, we recognize that national samples are not simple aggregations of individuals from an IID population, but combinations of subclasses from subpopulations with diverse characteristics. The composition of national populations from diverse domains deserves attention, and it also serves as an example for the two types of combinations that follow. Furthermore, these remarks are pertinent to combinations not only of national samples but also of cities, institutions, establishments, *etc.*

In recent years two kinds of sample designs have emerged that demand efforts beyond those of simple national samples: a) periodic samples and b) multipopulation designs. Each of these has emerged only recently, because they had to await the emergence of three kinds of resources: 1. effective demand supported by financial and political resources; 2. adequate institutional technical resources in national statistical offices; 3. new methods. In both types of designs we should distinguish the needs of the *survey methods* (definitions, variables, measurements), which must be harmonized, standardized, from *sample designs*, which can be designed freely to fit national (even provincial) situations, provided they are probability designs

(Kish 1994). Both types have been designed first and chiefly for comparisons: periodic comparisons and multinational comparisons, respectively. But new uses have also emerged: "rolling samples" and multinational cumulations, respectively. Each type of cumulation has encountered considerable opposition, and needs a new outlook, a new paradigm.

"Rolling samples" have been used a few times for local situations (Mooney 1956; Kish, Lovejoy and Rackow 1961). Then they have been proposed several times for national annual samples and as a possible replacement for decennial censuses (Kish 1981, 1990). They are now being introduced for national sample censuses first and foremost by the US Census Bureau (Alexander 1999; Kish 1990). Recommending this new method, I have usually experienced opposition to the concept of averaging periodic samples: "How can you average samples when these vary between periods?" In my contrary view, the greater the variability the less you should rely on a single period, whether the variation is monotonic, or cyclical, or haphazard. Hence I note two contrasting outlooks, or paradigms. Quite often, the opposition disappears after two days of discussion and cogitation.

"For example, annual income is a readily accepted aggregation, and not only for steady incomes but also for occupations with high variations (seasonal or irregular). Averaging weekly samples for annual statistics will prove more easily acceptable than decennial averaging. Nevertheless, many investors in mutual stock funds prefer to rely more on their ten-year or five-year average earnings (despite their obsolescence) than on their up-to-date prior year's earnings (with their risky "random" variations). Most people planning a picnic would also prefer a 50 year average "normal" temperature to last year's exact temperature. There are many similar examples of sophisticated averaging over long periods by the "naïve" public. That public, and policy makers, would also learn fast about rolling samples, given a chance."

(Kish 1998)

Like rolling samples, combining multipopulation samples also encountered opposition: national boundaries denote different historical stages of development, different laws, languages, cultures, customs, religions, behaviors. How then can you combine them? However, we often find uses and meanings for continental averages; such as European birth and death rates, or South American, or sub-Saharan, or West African rates. Sometimes even world birth, death, and growth rates. Because they have not been discussed, they all usually combined very poorly. But with more adequate theory, they can be combined better (Kish 1999). But first the need must be recognized with a new

paradigm for multinational combinations, followed by developing new and more appropriate methods.

5. EXPECTATION SAMPLING

Probability sampling assures for each element in the population ($i = 1, 2, \dots, N$) a known positive probability ($P_i > 0$) of selection. The assurance requires some mechanical procedure of chance selection, rather than only assumptions, beliefs, or models about probability distributions. The randomizing procedure requires a practical physical operation that is closely (or exactly) congruent with the probability model (Kish 1965). Something like this statement appears in most textbooks on survey sampling, and I still believe it all. However, there are two questionable and bothersome objections to this definition and its requirements.

The more important of the two objections concerns the frequent practical situations when we face a choice between probability sampling and expectation sampling. These occur often when the easy, practical selection rate for listing units of $1/F$ yields not only the unique probability $1/F$ for elements, but also some with variable k_i/F for the i th element ($i = 1, 2, \dots, N$) and with $k_i > 0$. Examples of $k_i > 1$, usually a small integer, occur with duplicate or replicate lists, dual or multiple frames of selection, second homes for households, mobile populations and nomads, farm operators with multiple lots. Examples of $k_i < 1$ are selecting a single adult from households, selecting single dwellings from buildings. In these examples often the k_i can be easily ascertained, and it is cheaper, more convenient and economical to use weighting than attempting to obtain $1/F$ for all the elements. These problems are described in books and articles.

In most cases, we find it more convenient and less expensive to accept the variable probabilities and to counter them with weighting the expected values $1/k_i$ or k_i than to operate another stage of selection. Thus, to paraphrase probability sampling: *expectation sampling* assures for each element in the population ($i = 1, 2, \dots, N$) a known positive expected number of selections ($k_i/F > 0$). These procedures are used in practice for descriptive (first order) statistics where the k_i or $1/k_i$ are neither large nor frequent. The treatments for inferential – second order or higher – statistics are more difficult and diverse, and are treated separately in the literature. Note that probability sampling is the special (and often desired) situation when all k_i are 1.

The other objection to the term probability sampling is more theoretical and philosophical and concerns the word "known" in its definition. That word seems to imply belief. Authors from classics like John Venn and M.G. Kendall to modern Bayesians like Dennis Lindley – and beyond at both ends – have clearly assigned "probability" to states of belief and "chance" to frequencies generated by objective phenomena and mechanical operations. Thus, our insistence

on operations, like random number generators, should imply the term "chance sampling." However, I have not observed its use and it also could lead to a philosophical problem: the proper use of good tables of random numbers implies beliefs in their "known" probabilities. I have spent only a modest amount of time on these problems and agreeable discussions with only a few colleagues, who did agree. I would be grateful for further discussions, suggestions and corrections.

6. SOME RELATED TOPICS

We called for recognition of new paradigms in six aspects of survey sampling, beginning with statistics itself. Finally, we note here the contrast of sampling to other related methods. Survey methods include the choice and definition of variables, methods of measurements or observations, control of quality (Kish 1994; Groves 1989).

Survey sampling has been viewed as a method that competes with censuses (annual or decennial), hence also with registers (Kish 1990). In some other context, survey sampling competes with or supplements experiments and controlled observations, and clinical trials. These contrasts also need broader comprehensive views (Kish 1987, section A.1). However, those discussions would take us well beyond our present limits.

REFERENCES

- ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Bulletin of the International Statistical Institute*, Helsinki, 52nd session.
- FISHER, R.A. (1926). *Statistical Methods for Research Workers*. London: Oliver & Boyd.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953a). *Sample Survey Methods and Theory, I – Methods and Applications*, New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953b). *Methods and Applications, II – Theory*. New York: John Wiley & Sons, Inc.
- JENSEN, A. (1926). The representative method in practice, *Bulletin of the International Statistical Institute*, 22, pt. 1, 359-439.
- KENDALL, M.G. (1968). Chance. *Dictionary of the History of Ideas*, (Ed. P.P. Wiener), New York: Chas Scribners.
- KIAER, A.W. (1895). The Representative Method of Statistical Surveys, English translation, 1976, Oslo: Statistik Sentralbyro.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1981). *Using Cumulated Rolling Samples*. Washington DC: Library of Congress.
- KISH, L. (1985). Chance, statistics, sampling. *Journal of Official Statistics*, 1, 35-47.
- KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.
- KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-79.
- KISH, L. (1994). Multipopulation survey designs. *International Statistical Review*, 62, 167-186.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.
- KISH, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25, 129-138.
- KISH, L., LOVEJOY, W. and RACKOW, P. (1961). A multistage probability sample for continuous traffic surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 227-230.
- KRUSKAL, W.H., and MOSTELLER, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review*, 47, 13-24.
- KRUSKAL, W.H., and MOSTELLER, F. (1979b). Representative sampling, II: Non-scientific literature. *International Statistical Review*, 47, 111-127.
- KRUSKAL, W.H., and MOSTELLER, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review*, 47, 245-265.
- KRUSKAL, W.H., and MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics. *International Statistical Review*, 48, 169-195.
- MOONEY, H.W. (1956). *Methodology of Two California Health Surveys*, US Public Health Monograph 70, Washington DC: US Government Printing Office.
- NEYMAN, J. (1934). On the different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- O'MUIRCHARTAIGH, C., and Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: a review. *Bulletin of International Statistical Institute, 43rd Session*, 1, 465-493.
- PORTER, T.M. (1987). *The Rise of Statistical Thinking: 1820-1900*, Princeton, NJ: Princeton University Press.
- STIGLER, S.M. (1986). *History of Statistics*, Cambridge: Harvard University Press.
- UNITED NATIONS STATISTICAL OFFICE (1950). *The Preparation of Sample Survey Reports*, New York: UN Series C No 1; also Revision 2 in 1964.
- VON MISES, R. (1939). *Probability, Statistics, and Truth*, London: Wm. Hodge and Co.

Still Rolling: Leslie Kish's "Rolling Samples" and the American Community Survey

CHARLES H. ALEXANDER¹

ABSTRACT

Leslie Kish long advocated a "rolling sample" design, with non-overlapping monthly panels which can be cumulated over different lengths of time for domains of different sizes. This enables a single survey to serve multiple purposes. The Census Bureau's new American Community Survey (ACS) uses such a rolling sample design, with annual averages to measure change at the state level, and three-year or five-year moving averages to describe progressively smaller domains. This paper traces Kish's influence on the development of the American Community Survey, and discusses some practical methodological issues that had to be addressed in implementing the design.

KEY WORDS: Rolling sample; Multi-year averages; Asymmetrical cumulations.

1. INTRODUCTION

A "rolling sample design", defined below, gives a single survey the flexibility to serve multiple purposes. The concept was developed by Leslie Kish in a series of papers (including Kish 1979a, 1979b, 1981, 1983, 1986, 1990, 1997, 1998 and Kish and Verma 1983, 1986) in which he elaborated the principles of cumulating information over space and time from a rolling sample. Kish advocated its use for a variety of purposes (Kish 1998), especially in developing countries (Kish 1979b), but also in the context of the U.S. census (Kish 1981). His personal use of rolling samples goes back at least to 1958, under the name "continuous sampling" (Kish, Lovejoy and Rackow 1961); a still earlier project (Mooney 1956) is cited in Kish (1998).

The American Community Survey (ACS), which is being developed as a replacement for the traditional "long form" survey conducted as part of the census, will use a form of the rolling sample design. This paper describes how the rolling sample concept is being implemented for the ACS, influenced by its specific objectives and operational considerations. The design decisions made for the ACS illustrate some issues that may arise for rolling samples in general. They also illustrate how Leslie Kish influenced survey development on multiple levels: philosophical, personal, and practical.

2. ROLLING SAMPLES

A "rolling sample" design jointly selects k non-overlapping probability samples (panels), each of which constitutes $1/F$ of the entire population. One panel is interviewed each time period until all the sample has been interviewed after k periods. Depending on the precision requirements, a single panel of $1/F$ may be sufficient to provide good estimates for the population as a whole, and

possibly for some large domains. For smaller domains or for greater precision for large domains, cumulations of different numbers of consecutive panels can be used, up to k/F of the population. A rolling sample design with $k=F$ is called a "rolling census". For a monthly rolling sample, it is natural to have F be a multiple of twelve, and natural cumulations are quarterly, semi-annual, annual, and multiple years.

"Domains" include both geographic areas and demographic subgroups. Kish (1987, section 2.3) presents a framework for the tradeoff between geographic and demographic detail, for a given required level of precision. Even more central to the idea of rolling samples was the idea of "asymmetrical cumulation" of data, over different lengths of time for different sizes of domain (Kish 1990, 1998), which was later broadened into a view of the basic similarities of averaging over space and averaging over time (Kish 1998), as well as averaging over different demographic domains. The flexibility of the rolling sample design comes from the opportunities it provides to make different tradeoffs between spatial, temporal, and demographic detail.

Leslie Kish left his colleagues with a challenge to extend these ideas into a "theory of combining populations" (Kish 1999, 2001). He organized a contributed paper session on "combining surveys" at the 1999 meetings of the International Statistical Institute, explaining to the presenters that we were all working on different aspects of the same problem, whether we knew it or not. The scope of this problem includes various forms of cumulation of data from rolling samples, as well as the question of how to combine data from different countries into statistics for larger entities such as the European Union. Kish (2001) suggests that these problems have fundamental features in common with the problem of combining information from different experiments (Cochran 1937, 1954).

¹ Charles H. Alexander, U.S. Bureau of the Census, Suitland, Maryland, U.S.A. 20233.

3. THE CENSUS LONG FORM AND INTERCENSAL ALTERNATIVES

The decennial census "long form" survey is the main source of subnational data about the *characteristics* of the U.S. population and housing. Estimates of the *number* of people and housing units come from the "short form" part of census administered to all households. With an overall sampling rate of one-in-six, the long form survey provides precise, detailed ("Precise" refers to the sampling error, and "detailed" means that estimates are given for many demographic domains within the geographic domain.) estimates of a variety of demographic and economic characteristics for individual states, large cities, and large counties or groups of counties. It provides useful, though less precise and less detailed, estimates for even very small areas such as small towns and Indian Reservations, as well as census tracts, which average about 4,000 population. For the smallest governmental units, higher sampling rates are used, as high as one-in-two for the smallest places, so that there are usable estimates for these areas. To compensate for the higher sampling rates in these areas, the rate is one-in-eight in the largest tracts.

Between the censuses, the federal government's statistical programs provide relatively little information about the characteristics of the population below the national level. The basic census counts are updated by an intercensal demographic estimates program, but other demographic and economic characteristics are available mainly from national surveys. The Current Population Survey (CPS), the U.S. monthly labor force survey, has about a one-in-1000 sampling rate with substantial overlap in the sample units from one month to the next so that the sample cannot be profitably cumulated over time as a rolling sample can. A March Supplement to the CPS collects additional information once a year, providing estimates for income and poverty at the state level, but with limited precision and demographic detail. There are programs which use modeling methods based on administrative records to make small-area estimates for unemployment, and for income and poverty, but not for a variety of characteristics.

The need for more frequent information for smaller domains (or "communities") has long been recognized (Hauser 1942; Eckler 1972, page 212; Bounpane 1986). Leslie gave credit to his friend, Philip Hauser, for proposing an "annual sample census" in 1941. Kish (1981) proposed a rolling sample as a way to meet this need, presenting several options including a rolling sample for the CPS. Instead a mid-decade census was authorized for 1985, but it was never funded. Nor was a proposal to double the size of the CPS (Tupek, Waite and Cahoon 1990).

Interest at the Census Bureau in intercensal information about population characteristics was revived by a proposal for a "Decade Census Program" advanced by Herriot, Bateman and McCarthy (1989). This program would have collected data in different states in different years;

ultimately this proposal did not gain acceptance. However, Roger Herriot's energetic and eloquent advocacy of the importance and potential value of intercensal subnational data created awareness in federal statistical agencies of the possibility of a "new paradigm" for the decennial cycle of data collection. Awareness of Kish's rolling sample proposal was definitely a factor during this period, as the Bureau considered new approaches for the 2000 census (see Bounpane 1986).

There was renewed Congressional interest in intercensal characteristics data (Melnick 1991; Sawyer 1993), and a "continuous measurement" alternative to the census long form was considered as part of the research for Census 2000, starting in 1992. Kish's rolling sample design was eventually proposed for this purpose because it provided flexibility in making estimates, as well as the potential for efficient data collection (Alexander 1993, 1997; National Academy of Sciences 1994, 1995). My recollection is that the most influential articles were Kish (1981, 1990), and that Kish and Verma (1983, 1986) were also consulted. "Continuous Measurement" was later renamed the "American Community Survey (ACS)".

The proposed ACS was not adopted for Census 2000, but after limited testing during 1996-1998, the ACS methodology was implemented in 36 counties for the years 1999-2001, so that ACS results could be compared to the 2000 census long form data. There was also a large-scale test in 2000, for a state-representative annual sample of about 700,000 addresses called the Census 2000 Supplementary Survey, of collecting long-form data separately from the census, using the ACS questionnaire. In 2001 and 2002, the Supplementary Survey is being continued, as part of the transition to the ACS.

4. THE PLANNED AMERICAN COMMUNITY SURVEY

The ACS will start in 2003, if funded by Congress, with a monthly sample of about 250,000 addresses, a new panel of addresses starting each month. This corresponds to a monthly rolling sample with an average rate of approximately $F = 480$ or an annual sample with $F = 40$. The survey will use $k = 60$, with the shortest published cumulation being calendar year estimates. The ACS will be conducted by mail, with nonresponse followup by telephone. A random sample of one-third of the remaining nonrespondents will be selected for followup in person.

For domains with average response rates, with a monthly $F = 480$, the standard errors for a 5-year average estimate from the ACS will be somewhat larger than for a corresponding estimate from the census long form, typically on the order of 1.33 times as large. This was judged to be "sufficiently close" for most purposes, given the advantage of timeliness and the expected lower missing data rates due to having a permanent staff of interviewers. In areas with

lower-than-average mail response rates, the subsampling for nonresponse follow-up will reduce the effective sample size. This happens not only because the number of interviews is reduced, but also because the unequal weights typically lead to a higher design effect (Kish 1965, pages 429-431). To compensate for this, the ACS will probably use a higher nonresponse subsampling rate in low-response areas, balanced by a lower sampling rate in areas with higher-than-average mail response. The details of this are still being determined. There also will be an oversample of addresses in small governmental units, as with the census long form sample.

An important development in the last decade, that made the ACS possible, (Kish (1981) suggests an alternative approach of "cumulative rolling listings", but this would be quite expensive for making regular estimates for all of the smallest domains, such as census tracts.) is the Census Bureau's program to maintain an ongoing Master Address File (MAF), linked to our TIGER geographic database. The main source of address updates throughout the decade is the Postal Service's Delivery Sequence File (DSF). The Bureau is implementing a MAF/TIGER modernization program that will augment the DSF updates with new addresses from data files provided by local governments, and from other administrative sources. This will be supplemented by new addresses encountered by interviewers from the ACS and other surveys in more rural areas. The monthly samples are actually generated by selecting an annual sample from the MAF in the previous September, and dividing it into 12 monthly panels. In February, there is a supplemental sample of new units from the DSF, spread across the remaining months of the year.

Replacing the 2010 census long form, by the ACS, is one component of a program to re-engineer the 2010 census. This also includes the modernization of MAF/TIGER, as well as a program of early research and testing to automate, streamline, and improve the census operations for 2010. This combination of improvements is expected to have a budgetary cost for the full 10-year cycle that is less than the cost of repeating the Census 2000 methods in 2010. This is a quite different plan than the vision of ACS described in National Academy of Sciences (1994, Chapter 6; 1995, Chapter 6), where I expressed hopes that eliminating the long form by itself, without other fundamental improvements, might save enough to pay for the ACS.

5. SOME VARIATIONS ON THE BASIC DESIGN, AND SOME ISSUES

5.1 Multi-stage Cluster Samples

The ACS uses an unclustered one-stage systematic sample, because the goals include providing data for all small geographic domains, such as tracts or block groups, each year. From discussions in Kish (1981, 1998), it is clear that rolling samples can also use cluster samples and

multiple stages of selection, as well as varying probabilities of selection. However, to qualify as a "rolling sample", the primary sampling units themselves must be a rolling sample. A design with a fixed set of primary sampling units (PSUs), with a rolling sample within each PSU, is a "cumulated representative sample" (Kish 1998).

Leslie was emphatic that the proposal by Herriot *et al.* (1989), was not what he meant by "rolling sample". However, it would seem to fit the definition as stated in Section 2, if the states are considered as PSUs. I think this demonstrates that there is an implicit requirement that a rolling sample must yield a useful representative probability sample in each time period, for each geographic domain of interest; this additional requirement does not hold if the PSUs are states. This requirement means that the clusters or PSUs need to be substantially smaller than the smallest domain of interest. (See Kish 1998, page 38.)

5.2 Differential Sampling Rates

Kish (1998, section 4) notes that a rolling sample can use different sampling fractions in different strata. This can get complicated, especially if the sampling fractions change over time, because the conditional probability of selecting a unit (without replacement) for the j^{th} panel in the h^{th} stratum depends on the sampling rates used in the previous panels in that stratum. This is even more complicated if the strata change over time, for example as the boundaries of governmental units change.

To simplify this for the ACS, we select the sample in two stages. The first stage selects a rolling "super sample" using a constant sampling rate for each panel and each year, equal to the largest sampling rate required in any stratum. The second stage subsamples the initial sample, to give the desired sampling rate for each stratum for that year. The selection of subsequent samples, which avoids overlap with the entire previous supersamples, needs only to keep track of the sampling rate for the first stage.

5.3 Updates to the Frame

In practice, the population is a little different for each panel. New addresses are added to the frame. Some old addresses cease to exist; they may be removed from the address list, or they may stay on the list and be deleted only after attempts to contact them. This presents no fundamental conceptual problem. It does mean that a "rolling census" would not necessarily contact every population unit that ever exists, since some units may go in and out of existence too quickly to fall into sample.

To avoid record-keeping of different conditional sampling rates for different "cohorts" of addresses which were added during Master Address File updates at different times, we have found it convenient to assign artificial "back samples" by selecting addresses from each set of new addresses not only for the current panel, but for past panels. These units are not interviewed, since the times for their assigned panels are past, but they are avoided during the without-replacement selection of future panels.

5.4 What Happens After Panel k ?

One question Leslie did not address explicitly, as far as I know, is how to draw the sample for panel $k+1$. I think he assumed that panel $k+1$ would be the same as panel 1, panel $k+2$ repeats panel 2, and so forth. This works fine for a simple random sample, but not so well for a systematic sample intended to spread the sample over a geographically sorted list, because as the frame changes over time, panel 1 doesn't keep its even spacing.

Our plan is to select panel $k+1$, and future panels, as a fresh systematic sample. Each one will avoid overlap with the previous $k-1$ panels, so there will always be k consecutive non-overlapping panels, but we won't worry about overlapping with panels before that.

5.5 Questionnaire Reference Date, Given an Extended Interview Period

The interviews from each monthly ACS panel take place over a three-month period, allowing two months for mail returns and telephone followup before starting the more expensive personal visits in the third month. Thus, the data actually collected in June consist of early mail returns from the June panel, late mail returns and telephone interviews from the May panel, and personal-visit followup cases from the April panel. This raises the issue of whether to ask the survey questions as of the time the survey was mailed out – the best choice as far as sampling bias – or as of the time the questions are asked – the best choice as far as response error and other nonsampling errors, especially for people who have moved from the address.

Taking these quality tradeoffs into account, we chose to use a "current" reference date, collecting the characteristics of the household members at the time of interview. One reason for this decision is that we think the nonsampling errors will be harder to evaluate than the sampling bias. Also the sampling biases in the monthly estimates will tend to cancel over the course of the year. This is one reason for limiting the ACS to annual and multiple-year estimates.

5.6 Use of Intercensal Population Estimates as Survey Weighting Controls

The Census Bureau has a program of "intercensal" (Leslie would call these "post-censal" estimates, reserving "intercensal" for estimates between two censuses that have been completed.) demographic estimates, based on demographic models. These models update the previous census, using vital records and other administrative records information. These estimates are used as independent weighting controls, or "post-stratification factors", for most national household surveys (see Kish 1965, pages 90-92). Adjusting the survey weights to agree with controls can reduce the variances of survey estimates, adjust for differences in coverage by age, sex, race, or Hispanic origin, and improve consistency across surveys. The census long form similarly uses the census counts as controls in its weighting.

The weighting controls have traditionally not been available for the smallest geographic domains, at least not with the demographic detail available for larger areas. Plans to produce more detailed controls for use in ACS weighting are described in Alexander and Wetrogan (2000). Some improvements will come from improved sources of administrative data, but in addition the ACS itself will provide information on changes in the population, which can be incorporated into the demographic models. The problem is complicated by the differences between the "current resident rule" used in the ACS and the "usual resident rule" used in the census; the ACS includes a question about part-year residents to help in adjusting for this difference. To facilitate this integration of survey data and demographic models, and especially to develop error measures for the resulting estimates, the Census Bureau is trying to develop "statistical" versions of the demographic models used in producing the intercensal population estimates. The inspiration for this effort to blend the statistical and demographic approaches is Purcell and Kish (1979).

6. DIFFERENT CUMULATIONS FOR DIFFERENT PURPOSES

For the main ACS objective, to replace the census long form as a source of detailed descriptive statistics, we plan to use 5-year ACS cumulations, for a data product similar to traditional long form "summary files". This is the shortest time period for which the ACS sampling error is judged to be reasonably close to that of the census long form. All sizes and types of geographic areas would be included on these 5-year data files. For allocating government funds based on an assessment of current need for the funds, simulations suggest that 3-year cumulations may be preferable to the 5-year, sacrificing precision for greater recency (Alexander 1998).

For individual areas, the most prominently published data will be one-year averages for areas greater than 65,000 population, and 3-year averages for areas greater than 20,000, in addition to the 5-year averages for all areas. Annual average estimates for areas below these thresholds will be available for more "sophisticated" uses to use in time series models, and to indicate large variations within the multi-year averages, but will not be as prominently displayed in our publications or on our websites.

These planned published ACS data products are designed to encourage analysts to use the same length of cumulation when comparing areas of different sizes, on the grounds that to do otherwise may be perceived as unfair to smaller jurisdictions. In doing this, we have accepted the notion of "asymmetrical cumulations" as far as levels of geography, but not necessarily within the same level of geography. For example, we would use one year for comparing states, but would recommend 5-years for all the

counties in a table comparing large and small counties. In this latter recommendation, we differ somewhat from Kish (1998, pages 42-43) which would let us use tables of counties with one-year estimates for large counties, 3-year averages for medium-sized ones, and 5-year averages for small ones. It will be interesting to see what practices data users will adopt in this regard.

7. WEIGHTING THE YEARS IN MULTI-YEAR CUMULATIONS

Kish (1998) points out that there are a number of choices for weighting multi-year cumulations. If there are ten yearly means \bar{y}_i , then there are many choices of $\bar{y} = \sum w_i \bar{y}_i$, with $\sum w_i = 1$, to use as the ten-year cumulations.

For the ACS 5-year and other multi-year cumulations, discussed in section 6, our plans are to give the years equal weights in the standard published data products, e.g., $w_i = 0.2$ for the 5-year average. This was an area of disagreement with Kish (1998), which gently urges us to consider of alternatives, in particular weights of the form $w_{i+1} = Cw_i$, with $C > 1$.

An underlying issue in thinking about unequal weights is what statistical problem we are trying to solve. Using the 2003 – 2007 cumulation as an example, is the goal:

- to provide a “direct design based” estimate for the 2003 – 2007 historical average;
- to provide a “model-based” estimate for the 2007 value; or
- to provide a “direct, design-based” estimate for a weighted 2003 – 2007 historical average, with more weight on recent years?

To interpret the 2003 – 2007 estimate as an estimate for 2007 requires a model or assumptions about the time series for the area. The problem may be viewed as combining a direct estimate for 2007 with a forecast for 2007 based on the years 2003, ..., 2006, with the requirement that the same formula be used for all areas and all characteristics to preserve additivity in the tables and comparability across tables.

I have previously interpreted the decision as a choice between the first two goals, and have shied away from the second approach for the ACS, ultimately because of the concerns expressed in Hansen, Madow and Tepping (1983, sections 3 and 5.5) about using model-based estimates for general-purpose “official statistics”. With the variety of statistics and geographic areas covered by the ACS, there inevitably will be some where the compromise model fails badly; a data user may be unaware of this failure, or may be very aware. In what sense can the compromise average be viewed as a valid estimate for 2007 when the compromise model clearly fails, and what measure of error would be associated with it? With this view of the issue, we have

recommended using the unweighted multi-year averages as the standard general-purpose data product, with the time series of annual estimates being available for use in time series models for specific applications, and for interpreting the multi-year averages when there is variation within the 5-year period.

However, upon rereading Kish (1998), I now interpret his view of the weighted average to be the third formulation, a design-based estimator of a more up-to-date population parameter. This avoids the concerns about model fit for general-purpose uses, although there is still the question of how to justify and achieve a consensus solution. Also, the unequal weights tend to increase the standard errors of the multi-year averages. But Kish (1998, page 40) will get the last word on the subject:

“Important questions remain for further discussions and research. Perhaps forever, and this can become a ‘growth industry.’”

8. NOT COMBINING THE CPS AND THE ACS

Leslie often said he was pleased to see his idea being implemented in the ACS, but I think he was disappointed that we did not try to replace both the census long form and the CPS with one survey. By contrast with some other issues where we had lively discussions, Leslie took a “hands off” stance on this issue. I think he viewed this as a decision about quality tradeoffs, which the government agencies had to work out for ourselves. There were two main reasons for our decision:

We cannot adequately measure the monthly unemployment rate with a mail survey. Correct measurement of the unemployment rate requires complex questions that would not be feasible to ask by mail, for example, to probe to be sure that someone who is “looking for work” did conduct an active job search. (See Butani, Alexander and Esposito 1999). The Census 2000 Supplementary Survey, using the ACS procedures, dramatically overestimated the 2000 national unemployment rate (5.3 percent versus 4.0 percent in the CPS). A similar difference was seen in the 1990 census.

A mail survey would lag substantially in producing monthly rates, compared to the CPS. In addition, the impossibility of completing all the mail interviews for a panel in the designated month introduces biases in monthly estimates (see section 5.5 above). These problems would be reduced somewhat for quarterly moving averages instead of monthly estimates, which Leslie frequently suggested (for example Kish 1999), but the monthly unemployment report is an indispensable economic indicator in the U.S.

It is too expensive to replace the long form without using mail. A rolling sample survey, conducted in person with a large enough sample to replace the long form, would have to be 3 or 4 times as large as the CPS. This is a function of

the size of the U.S. population, and the number of tract-sized domains for which estimates are required from the long form. Such a survey would be much more expensive per case than the CPS, because it could not use a cluster sample or telephone interviews for repeated interviews of the same households, as does the CPS. The total cost of such a survey would be several times as great as the combined cost of the proposed ACS and the CPS.

Because it is designed so narrowly as a long form replacement, the ACS does not illustrate the full range of flexibility that Leslie envisioned from a rolling sample. Under different circumstances, for a smaller population, with less need for very small domains from the "long form survey", or less strict requirements for timing and questions for the labor force survey, it might be possible for a labor force survey with a rolling sample to meet the demands for small domain data. With the further addition of a split panel or other components (Kish 1998, pages 40-41) an even wider range of objectives could be met.

9. CONTRIBUTIONS: PHILOSOPHICAL, PERSONAL, AND PRACTICAL

The long list of articles by Leslie Kish on the subject of rolling samples clearly demonstrates the intensity and tenacity of his campaign for what he understood as an important idea. The evolution of the idea over the course of these papers also illustrates the depth of his attention to "philosophical" questions about the fundamental quality objectives for a survey: What are we trying to do? How does the choice of survey design relate to what we are trying to do, and why? This kind of guidance is crucial at the start of a survey program, when the "big questions" are being addressed, and makes the difference between ideas that quickly fall by the wayside and those that are "still rolling".

Leslie's personal support of other statisticians went far beyond his papers. Though I was by no means one of his closest colleagues, he regularly provided personal advice or encouragement when he sensed it was needed. The "still rolling" in this paper's title was the title I used in e-mail messages to him when I had news about the ACS's perilous passage through the annual budget cycle, which was most of the time. He would respond briefly by e-mail, but important messages always came in the form of handwritten letters.

Finally, based on these papers, it is clear that Leslie was always a practical person, even at his most philosophical, and that his papers cannot be fully appreciated without knowing what was going on in the survey world when he wrote them. Looking back over his rolling sample papers, I can see many comments, about both details and general principles, that were aimed at enlightening specific decisions that the Census Bureau needed to make at the time. I would guess that throughout his work, there are

specific messages to help out someone somewhere in the world who faced a practical design decision at the time.

REFERENCES

- ALEXANDER, C.H. (1993). A continuous measurement alternative for the U.S. Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 486-491.
- ALEXANDER, C.H. (1997). The american community survey: Design issues and initial test results. *Proceedings of Symposium 97, New Directions in Surveys and Censuses*, 187-192.
- ALEXANDER, C.H. (1998). Recent developments in the american community survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 92-100.
- ALEXANDER, C.H., and WETROGAN, S. (2000). Integrating the american community survey and the intercensal demographic estimates program. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 295-300.
- BOUNPANE, P. (1986). How increased automation will improve the 1990 census. *Journal of Official Statistics*, 4, 545-553.
- BUTANI, S., ALEXANDER, C. and ESPOSITO, J. (1999). Using the american community survey to enhance the current population survey: Opportunities and issues. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Statistical Policy Working Paper 29*, 3, 102-111.
- COCHRAN, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (B)*, 4, 102-118.
- COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- ECKLER, A.R. (1972). *The Bureau of the Census*. New York: Praeger Publishers.
- HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 384, 776-793.
- HAUSER, P.M. (1942). Proposed annual census of the population. *Journal of the American Statistical Association*, 37, 81-88.
- HERRIOT, R.A., BATEMAN, D.B. and MCCARTHY, W. F. (1989). The Decade Census Program - A new approach for meeting the nation's needs for sub-national data. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-355.
- KISH, L., LOVEJOY, W. and RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.
- KISH, L. (1979b). Rolling samples instead of censuses. *Asian and Pacific Census Forum*, G(1), August 1979, 1-2, 12-13.
- KISH, L. (1981). *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*. Washington, D.C., U.S. Government Printing Office.

- KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright). New York: Academic, 72-84.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 1-12.
- KISH, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons.
- KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-79.
- KISH, L. (1997). Periodic and rolling samples and censuses. Chapter 7 in *Statistics and Public Policy*, (Ed. Bruce D. Spencer). Clarendon Press, Oxford.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 1, 1998, 31-46.
- KISH, L. (1999). Combining/cumulating population surveys. *Survey Methodology*, 25, 2, 129-138.
- KISH, L. (2001). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, to appear in 2001.
- KISH, L., and VERMA, V. (1983). Census plus samples: Combined uses and designs. *Bulletin of the International Statistical Institute*, 50(1), 66-82.
- KISH, L., and VERMA, V. (1986). Complete Censuses and Samples. *Journal of Official Statistics*, 2, 381-93.
- MELNICK, D. (1991). The census of 2000 A. D. and beyond. *Reviews of Major Alternatives for the Census in the Year 2000*. U.S. Government Printing Office, Washington, D.C., August 1, 1991, 60-74.
- MOONEY, H.W. (1956). Methodology in two California Health Surveys. *U.S. Public Health Monograph*, 70.
- NATIONAL ACADEMY OF SCIENCES (1994). *Country People in the Information Age*. (Eds. D.L. Steffey and N.M. Bradburn). National Academy Press, Washington, D.C.
- NATIONAL ACADEMY OF SCIENCES (1995). *Modernizing the U.S. Census*. (Eds. B. Edmonston and C. Schultze). National Academy Press, Washington, D.C.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- SAWYER, T. C. (1993). Rethinking the census: Reconciling the demands for accuracy and precision in the 21st century. Presented at the research conference on undercounted ethnic populations, May 7, 1993.
- TUPEK, A. R., WAITE, P. J. and CAHOON, L. S. (1990). Sample expansion plans for the current population survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 72-77.

Redesign of the French Census of Population

JEAN-MICHEL DURR and JEAN DUMAIS¹

ABSTRACT

Census-taking by traditional methods is becoming more difficult. The possibility of cross-linking administrative files provides an attractive alternative to conducting periodic censuses (Laihonen 2000; Borchsenius 2000). This was proposed in a recent article by Nathan (2001). INSEE's redesign is based on the idea of a "continuous census," originally suggested by Kish (1981, 1990) and Horvitz (1986). A first approach that would be feasible in France can be found in Deville and Jacod (1996). This article reviews methodological developments since INSEE started its population census redesign program.

KEY WORDS: Balanced sampling; Census; Continuous census; Calibration.

1. INTRODUCTION

1.1 Reasons for the Redesign

France has been conducting censuses for many years to measure the *de jure* population of its administrative districts and to describe the socio-demographic characteristics of its territory at all levels of geography, from districts of communes to the country as a whole. The 1999 census was conducted in the usual manner: delivering and retrieving questionnaires by census interviewers, organisation, technical assistance and control by INSEE, execution by the Mayor as the state representative. For various reasons, however, we decided to re-examine the census.

First, the interval between censuses has a tendency to increase in length. Indeed, the periodicity of censuses is not covered by laws, and each census date is determined by a statutory order. Before the war, censuses were taken every five years; then the gap grew to seven years, then eight, the last census, originally planned for 1997, was postponed until 1999 for budgetary reasons, that is, 9 years after the previous census. Moreover, the public does not always understand the need for such a massive operation at a time when the number of administrative files is increasing, even though that same public has expressed serious concerns about the cross-referencing of such files. In addition, the decentralization that has been going on in France for over 20 years has generated numerous requirements for statistics in support of local policy-making. As the supreme source of local information, the census had to adapt to these changes and provide fresher yet still highly detailed data.

As a result, a population census redesign program was established at INSEE in the late 1990s. Since France has no population register and, in view of the circumstances, is unlikely to institute one, the decision was made to consider a compromise solution that would combine annual sample surveys with the use of non-nominative administrative files that INSEE is authorized to use solely for statistical

purposes. Communes whose population is below a certain threshold (10,000 for the moment) will be covered by annual take-all surveys with a rotation period of five years. For the other communes, a sample survey will be conducted each year, with the entirety of the commune being covered within the same five-year rotation period. To carry out this redesign, a new legal framework was needed. The project was submitted to the Conseil d'État, which recommended on July 2, 1998, that the government table draft legislation in Parliament.

Aside from the need to strengthen the census legal basis, the Conseil was of the view that since population counts were referred to in over 200 statutes or regulations, making a major change in the way they were produced would require legislative approval. Within this framework, the purpose of the legislation was essentially to set out the principles and rules governing the organization of the census.

The operation was placed under State responsibility and control: INSEE was to establish the collection framework (concepts, protocols), select the samples, ensure the quality of the information collected, and process and disseminate the data. The communes as local organisations, were to prepare and conduct the census surveys. The State would provide financial assistance to cover the costs. These arrangements clarify the role and responsibilities of each of the partners.

1.2 Quality Goals

The program has the following quality goals:

1.2.1 Data Quality

Timeliness: The goal is to be able to disseminate by the end of year A the *de jure* population of all administrative districts as at January 1 of year A-2; a statistical description of all geographic units (communes and commune groups, districts of major cities, lands, *etc.*) as of January 1 of year

¹ Jean-Michel Durr, Programme de rénovation du recensement de la population, INSEE, Direction générale, 18 boul. Adolphe Pinard, 75675 Paris CEDEX 14, France; Jean Dumais Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada. This paper was prepared while the author was on secondment at the Programme de rénovation du recensement de la population, INSEE.

A-2; and a statistical description of France and its major geographic units (regions, *etc.*) as of January 1 of year A. In comparison with the general census, the redesigned census will produce similar population and housing data an average of three to four years earlier.

Relevance: The data produced must be relevant to local needs. In particular, data that are worth studying only at levels of geography far above the commune will be set aside in favour of data that are more useful for local purposes. What data will be collected will be determined by the Conseil national de l'information statistique (CNIS), whose membership includes representatives of various categories of producers and users of public statistics. A CNIS working group has proposed changes while at the same time preserving the necessary continuity with previous censuses and limiting the response burden.

Precision: The census must provide data that are meaningful for all levels of geography in France. The data produced must be sufficiently precise, even at the sub-communal levels, for the most useful cross-tabulations at those levels. This means, in particular, distributions by sex and age, by type of activity and socio-professional category, and by type of housing. It must be possible to estimate the precision of the data, and users must be informed of that precision.

User-friendliness: To avoid annoying users, the data produced must be easy to understand and comparable in use to data produced by a general census.

1.2.2 Process Quality

Response burden: To limit the response burden for the public, the amount of information collected must be kept to a bare minimum. In particular, information available for the same level of geography from other sources will not be collected in the census unless it can be used to produce useful cross-tabulations with other variables. As in previous censuses, the personal questionnaire will be confined to one double-sided sheet of paper.

Questionnaire: Since collection is by the drop-off/pick-up method, the questionnaires must be universally accessible. To ensure that the questions will be understood, qualitative testing was conducted using focus groups. In addition, a collection test was carried out on 4,000 dwellings in the first half of 2001.

Confidentiality: Data gathered in the census are protected by law. Personal information collected in the census can be accessed only by authorized persons. The data are for INSEE and can be used only for statistical purposes. Only data essential to the preparation and conduct of census surveys are shared with communes or commune groups, on a need-to-know basis.

Quality of coverage: The coverage of general censuses was not systematically evaluated. Following the 1990 census, a postcensal survey indicated that the rate of undercoverage was about 1.8% and the rate of overcoverage was about 0.9%, for an overall precision of roughly 0.9%. The

largest undercounts were in large agglomerations. By conducting an annual sample survey in communes with a population of more than 10,000 and thereby reducing the number of people to be covered in the census, we will be able to focus our efforts on obtaining answers from respondents. The coverage of the redesigned census will be evaluated on a regular basis through comparison with administrative data and through special surveys.

Technical and organizational robustness: Because of the volume of data processed and the importance of the census, the program must be based on tried and true technical innovations. Furthermore, the robustness of the census apparatus must be evident in the launch of the operation. Technical or functional innovations can be introduced at any time in the census cycle as part of evolving maintenance or specific projects. The annual surveys can be used to test the effectiveness of such projects before they are applied to the entire process. However, major changes such as questionnaire updates will generally be made only for the beginning of a five-year cycle. The organization of the census will depend on a balanced partnership between INSEE and the communes. INSEE must be capable of building the proposed structure within its budget and its work program by reorganizing its operations. Similarly, the communes and intercommunal cooperation bodies must be able to support the census organization. The yearly cycle of surveying large communes and the option that small and medium communes will have of delegating collection to an intercommunal body are likely to promote the professionalization of collection workers.

With the integration of census operations into the annual work program of the regional offices, and the fact that the operation is one-seventh the size of the general census, INSEE will have tighter control of the census. Instead of having 110,000 census agents collecting data from 60 million people in 36,700 communes in a particular year, it will have only 18,000 agents visiting roughly 9 million residents in about 8,000 communes.

The division of responsibilities between INSEE and the communes, the resources that the communes will require, and the validation processes for the various stages will be set out in a decree.

Cost control: With the five-year collection cycle, the financial burden of conducting the census can be spread over a longer period. For communes with a population of more than 10,000, the cost of the redesigned census will be lower than the cost of the current census of population. On the other hand, for communes with fewer than 10,000 residents, the cost should be equal to that of a general census, but it would be every five years instead of the roughly eight-year cycle of the general census. The cost of the redesigned continuous census will be equivalent to one seventh of that of a general census. This will contribute to archive the reform without budget increase. However, a slightly larger budget in the first few years would help to iron the kinks out of the collection process.

2. SAMPLING STRATEGY

The commune is the linchpin of the redesign effort. The set of "small and medium-sized communes" (those with a population of less than 10,000) will be sampled at an average rate of 20% a year, and all their dwellings will be visited; all "large communes" will be visited annually, but only a fraction of their dwellings will be surveyed.

2.1 Small and Medium-sized Communes

Let's start with "small and medium-sized communes". In each region, five rotation groups of communes will be formed using data from the 1999 population census. They will consist of balanced samples (Deville and Tillé 1999, 2000) of the age-sex distribution of the communes' population. This approach should help minimize year-to-year variation due to sampling.

communes in Rhône-Alpes in the 1990 population census. For each rotation group, both the quartiles and the range of the distribution are shown. It is interesting to note how similar the charts are. The "number of women aged 20 to 39" variable was used to form the groups. Neither the number of principal residences nor any of the household or dwelling variables plays a part in the balancing.

Each year, the population and housing in all the communes in one rotation group will be fully enumerated. Hence, each "small and medium commune" will be completely enumerated once every five years, and a fifth of all the "small communes" will be covered each year.

2.2 Large Communes

The "large commune" sample will be based on the "répertoire d'immeubles localisés" (RIL) (inventory of located buildings). The RIL is a list of buildings (residential, institutional or commercial) identified individually so as to generate a digitized map. Initially, the RIL will be populated with data from the 1999 census, which will provide a statistical portrait of each residential building. (In the 1999 census, a building is defined as the set of dwellings served by the same staircase; thus, a single physical building can consist of more than one "census building".)

The RIL will be continually updated using building permits, demolition permits, utility records (water, gas, hydro, *etc.*), information supplied by local governments, and field observations. Thus, the RIL may be used to create a building sample frame for "large communes".

In each IRIS2000 (an IRIS2000 is a set of "îlots regroupés selon des indicateurs statistiques" (blocks grouped by statistical indicators), a homogeneous area with a population of about 2,000) of each "large commune", five rotation groups of addresses will be formed using the same sampling model as in "small and medium communes". Three additional strata will be created in each IRIS2000: one for industrial buildings (plants, warehouses, *etc.*), another for collective dwellings (institutions, group homes, communal groups, boarding schools, *etc.*) and a third for new addresses.

One fifth of the industrial buildings will be visited each year to verify that they contain no dwellings (custodian's quarters or space converted for habitation); any dwellings found in such buildings will be considered self-representing because of their special nature. All collective dwellings will be covered each year; 20% of them will be visited, and the population counts of the remaining 80% may be updated by telephone. Finally, all new residential buildings will be inserted in the rotation groups.

As noted above, each address rotation group will be visited once in each five-year period. A sub-sample of addresses, which corresponds to 40% of the dwellings of the group, will be selected. In each selected address, the complete dwelling content will be surveyed.

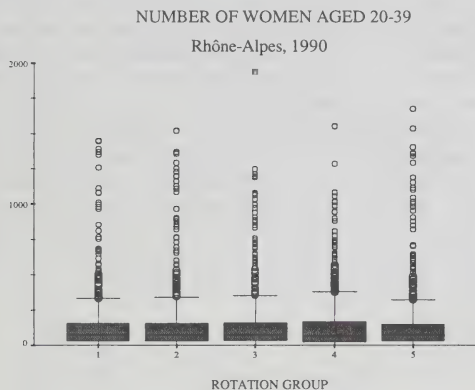


Figure 1

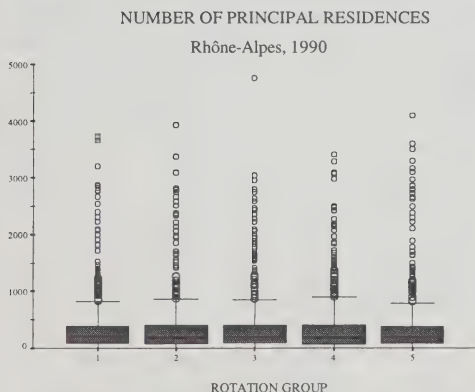


Figure 2

Figures 1 and 2 show how balanced the five rotation groups will be. They contain box-and-whisker diagrams of two variables measured in the 2,811 small and medium

$$\tilde{R}_{D,IV}^{A-2} = 0.2 \times \Theta_1 + 0.8 \times \Theta_2.$$

Similarly, for commune E in Group V, with Q1 and Q2 appropriately defined, we would have:

$$\tilde{R}_{E,V}^{A-2} = 0.4 \times \Theta_1 + 0.6 \times \Theta_2.$$

Adjustment factors Θ will have to be calculated for relatively detailed population strata, such as age-sex classes, so as to keep as much demographic and geographic flexibility as possible in the census adjustment. The quality of the administrative files and local disparities will dictate the level at which the adjustment can be made most conveniently (for départements, metropolitan areas, ...). The same process can be applied to large communes if we replace "small commune" with "address".

Finally, when every commune in every group has been imputed, the estimated total for a variable of interest from the imputed file (detailed estimates) is unlikely to match the total estimated from observations alone (overall estimates published two years earlier). It has therefore been decided that the detailed estimates will be calibrated on the overall estimates. Once again, the level of calibration will depend on local trends and the quality of the overall estimates.

3.3. De Jure Population Estimates

The de jure population estimates are the third set of estimates derived from the census. They are the population figures that are used, by law, to determine commune funding, electoral boundaries, the composition of municipal councils, *etc.*

The "total de jure population" of a commune includes persons

- whose principal residence is within the commune,
- who live in an institution or a collective dwelling located within the commune,
- who have a residence in the commune and live in an institution or a collective dwelling located in another commune but have kept a dwelling in their commune of origin,
- who live in a collective dwelling in another commune for work or live in another commune for education,
- who are attached to the commune for administrative purposes (itinerant workers, sailors and so on).

Clearly, these populations cannot be estimated until the entire territory of the commune has been covered, that is, until the detailed estimates have been produced.

3.4. Estimation of Sampling Variance

The global and detailed estimates will be accompanied by a measure of their statistical quality. Work on this project began in the fall of 2001. The preferred option at this time is to use reference tables, as is done in the Canadian Labour Force Survey, for example. The sampling

variances will probably be obtained by resampling the frame.

3.5. Imprecision Due to Synthesis

In the section 3.2, we showed how collected data will be used to produce synthetic estimates: first, an extrapolation for an "old" census, for two rotation groups (I and II, say); then directly using the census results for a third rotation group (III, say); and finally, combining extrapolations and backward projections to calibrate the last two groups (IV and V, say).

This synthesis can be formalized using a non-response model (Särndal 1990). The annual campaign is similar to a take-all survey that has an 80% non-response rate, which is dealt with using ratio imputation. If we let s represent the whole sample, r the respondents and $s-r$ the non-respondents, we have

$$y_k = \begin{cases} y_k & \text{if } k \in r \\ \hat{\beta} x_k & \text{if } k \in s-r \end{cases} \quad \text{with } \hat{\beta} = \frac{\bar{y}_r}{\bar{x}_r}.$$

Thus, the imputation model is

$$\xi: \begin{cases} y_k = \beta x_k + \epsilon_k \\ E(\epsilon_k) = 0 \\ V(\epsilon_k) = \sigma^2 x_k \end{cases}$$

where the errors ϵ_k are not correlated. With such a model, under simple random sampling,

$$\begin{aligned} \hat{Y} &= \frac{N}{n} \sum y_k = \frac{N}{n} \left\{ \sum_r y_k + \sum_{s-r} \hat{\beta} x_k \right\} = \dots \\ &= N \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s \end{aligned}$$

The uncertainty around estimation with imputation depends on the sampling errors and the quality of imputation model ξ :

$$\begin{aligned} (\hat{Y}_i - Y) &= (\hat{Y} - Y) + (\hat{Y}_i - \hat{Y}) \\ \text{Total} &= \text{sampling} + \text{imputation} \\ \text{uncertainty} &\quad \text{uncertainty} \quad \text{of model} \end{aligned}$$

If we assume that the imputation is unbiased:

$$E_\xi E_s E_r (\hat{Y} - Y) = 0$$

we have,

$$\begin{aligned} V_{\text{total}} &= E_\xi E_s E_r (\hat{Y} - Y)^2 = \dots \\ &= E_\xi E_s E_r (\hat{Y} - Y)^2 + E_\xi E_s E_r (\hat{Y} - \hat{Y})^2 \\ &= E_\xi V_s + E_s E_r V_\xi \\ V_{\text{total}} &= V_{\text{sample}} + V_{\text{imputation}} \end{aligned}$$

assuming that the design and response mechanism are independent from imputation. Using imputed data as if they were observed data to compute the estimate of V_s results in an underestimate of V_{sample} . In terms of expectation,

$$E_{\xi}(\hat{V}_s - \hat{V}_{\cdot s}) = V_{\text{dif}}.$$

For the estimators of these variances, Särndal shows that we get

$$\hat{V}_{\text{sampling}} = N^2 \left(\frac{1}{n} \frac{1}{N} \right) \{S^2 + C_0 \hat{\sigma}^2\}$$

with C_0 close to $\left(1 - \frac{m}{n}\right) \bar{x}_{s-r}$ and $\hat{\sigma}^2$ close to $\frac{\sum_r e_k^2}{\sum_r x_k}$ and

$$\hat{V}_{\text{imputation}} = N_2 \left(\frac{1}{m} - \frac{1}{n} \right) A \bar{x}_s \hat{\sigma}^2,$$

with $A = \bar{x}_{s-r} / \bar{x}_r$, which we can take as a respondent selection effect. We note that if $x_k = 1$, then we obtain a two-phase sample of size m in n and n in N . In addition, if $s = r$, $V_{\text{total}} = V_{\text{sampling}}$.

In Särndal's model, the x (administrative data) and y (census data) are contemporaneous; at the very least, we will have observed some of the y . Using the structure developed in the previous section, we would have:

Year A-2		
y_k	x_k	m respondents (Group III)
y_k	x_k	n-m imputations (other groups)

In the continuous census system, not everything is synchronous:

... A-4		A-3		A-2		A-1	A
Y_I^{A-4}	X_I^{A-4}		X_I^{A-3}		X_I^{A-2}		
	X_{II}^{A-4}	Y_{II}^{A-3}	X_{II}^{A-3}	Y_{II}^{A-2}	X_{II}^{A-2}		
	X_{III}^{A-4}		X_{III}^{A-3}	Y_{III}^{A-2}	X_{III}^{A-2}		
	X_{IV}^{A-4}		X_{IV}^{A-3}		X_{IV}^{A-2}
	X_V^{A-4}		X_V^{A-3}		X_V^{A-2}		...

That is, Y_{II}^{A-3} , X_{II}^{A-3} , Y_{II}^{A-2} , and X_{II}^{A-2} are not all measured or observed in the same year. In fact, if we look at Group III on its own, for example, we have a sample of size n in year A-2 and an identical but totally non-respondent sample in year A-3. Consequently, some parameters in the estimate of V_{total} cannot be calculated.

On the other hand, if we take the problem over a specific period, we have a sample of size n and $4n$ non-respondents. We could approximate the uncertainty of the asynchronous

imputation process (the process we have in the redesigned census) with the uncertainty of the synchronous imputation process (similar to Särndal's model).

This approach was tested on the small and medium communes of Rhône-Alpes, for which the rotation groups, 1990 property tax data and 1990 population census data are available (Kauffmann 2000). The method gives good results for variables that are highly correlated with property tax; the results also indicate that a source of administrative data that are similar to variables describing people will be necessary to maintain the model errors at an acceptable level.

4. WORK IN PROGRESS

The methodological work involved in redesigning the census is far from complete. The following projects are still under way:

- establishment of rules for crossing the size threshold, problems of oscillation around the 10,000 population threshold, and calculation of the de jure population;
- the sensitivity of stratum boundaries in large communes and their robustness over time;
- the updating and maintenance of sampling frames and samples, especially adjustments that may be required when a commune crosses the size threshold and the incorporation of new objects into rotation groups;
- massive imputation and synthesis, both models and their precision;
- estimation of the precision of estimators; and
- collecting data from mobile population groups.

REFERENCES

- BORCHSENIUS, L. (2000). From a Conventional to a Register-based Census of Population. Les Recensements après 2001, Séminaire Eurostat-INSEE, Paris.
- DEVILLE, J.C., and JACOD, M. (1996). Replacing the Traditional French Census by a Large Scale Continuous Population Survey. *Annual Research Conference Proceedings*, USBC, Washington.
- DEVILLE, J.C., and TILLÉ, Y. (1999). *Balanced Sampling by Means of the Cube Method*. CREST-ENSAI, working paper submitted for publication.
- DEVILLE, J.C., and TILLÉ, Y. (2000). Echantillonnage équilibré par la méthode du cube et estimation de variance. *Journées de Méthodologie*, December 2000, INSEE, Paris.
- HORVITZ, D.G. (1986). Statement to the Subcommittee on Census and Population. Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.

- KAUFFMANN, B. (2000). *Estimations annuelles dans la rénovation du recensement de la population*. Working paper, Département de la démographie, INSEE.
- KISH, L. (1981). Population Counts from Cumulated Samples. Congressional Research Service. *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.
- KISH, L. (1990). Rolling Samples and Censuses. *Survey Methodology* 16, 1, 63-71, Statistics Canada, Ottawa.
- LAIHONEN, A. (2000). 2001 Round Population Censuses in Europe. *Les Recensements après 2001*, Séminaire Eurostat-INSEE, Paris.
- NATHAN, G. (2001). Models for combining longitudinal data from administrative sources and panel surveys. Invited paper, ISI, Seoul, August 2001.
- SÄRNDAL, C.-E.(1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Ottawa, October 1990, 337-350.

Benchmarking Parameter Estimates in Logit Models of Binary Choice and Semiparametric Survival Models

IAN CAHILL and EDWARD J. CHEN¹

ABSTRACT

An approach to exploiting the data from multiple surveys and epochs by benchmarking the parameter estimates of logit models of binary choice and semiparametric survival models is developed. The goal is to exploit the relatively rich source of socio-economic covariates offered by Statistics Canada's Survey of Labour and Income Dynamics (SLID), and also the historical time-span of the Labour Force Survey (LFS), enhanced by following individuals through each interview in their six-month rotation. A demonstration of how the method can be applied is given, using the maternity leave module of the LifePaths dynamic microsimulation project at Statistics Canada. The choice of maternity leave over job separation is specified as a binary logit model, while the duration of leave is specified as a semiparametric proportional hazards survival model with covariates together with a baseline hazard permitted to change each month. Both models are initially estimated by maximum likelihood from pooled SLID data on maternity leaves beginning in the period 1993-1996, then benchmarked to annual estimates from the LFS 1976-1992. In the case of the logit model, the linear predictor is adjusted by a log-odds estimate from the LFS. For the survival model, a Kaplan-Meier estimator of the hazard function from the LFS is used to adjust the predicted hazard in the semiparametric model.

KEY WORDS: Microsimulation; Benchmarking; Semiparametric survival models; Binary logit.

1. INTRODUCTION

Researchers often base econometric models on a survey conducted over a short period of time. In this case it may be desirable to incorporate information from a supplementary data source covering a longer period, even if measurements are only available for the dependent variable. For a broad class of non-linear models, we develop a simple method of benchmarking the parameter estimates obtained from a survey rich in explanatory variables to information from a survey with significant historical depth. A primary objective is that model predictions accord with information from the secondary data source. We demonstrate application of the method first to a simple logit model of binary choice, and secondly to a semiparametric survival model. Since the survival model can be viewed as a sequence of binary choices, while retaining an interpretation as an incompletely observed continuous time model, it provides a natural generalization of the first application.

The illustration we provide is a study of maternity leave. The Statistics Canada Survey of Labour and Income Dynamics (SLID) provides data on both the incidence of choosing a maternity leave over withdrawing from the labour force, and on the duration of maternity leave, as well as a rich set of explanatory variables. Because of this we use SLID to estimate base parameters, including those determining the effects of the explanatory variables on the incidence (the logit model) and hazard of returning to work (the survival model). The Canadian Labour Force Survey (LFS) conducted by Statistics Canada provides reasonable proxies for both the incidence and duration extending back

to 1976. The SLID parameter estimates are therefore benchmarked to LFS estimates of incidence and the hazard of returning to work during the period 1976-1992, which is prior to the availability of SLID data.

The work was carried out while developing the maternity leave module of the LifePaths microsimulation model at Statistics Canada. The goal of the LifePaths project is to construct a dynamic microsimulation model encapsulating as much detail as possible on socio-economic processes in Canada, as well as the historical patterns of change in those processes. LifePaths has been employed in a broad range of policy analysis and research activities. Examples include Canada Student Loan policy (under contract to Human Resources Development Canada and the Government of Ontario), returns to education (Appleby, Boothby, Rouleau and Rowe 1999), time use (Wolfson and Rowe 1996; Wolfson 1997; Wolfson and Rowe 1998a), tax-transfer and pensions (Wolfson, Rowe, Gribble and Lin 1998; Wolfson and Rowe 1998b), and labour force careers (Rowe and Lin 1999). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1997; Rowe and Chen 1998; Plager and Chen 1999) and earnings correlation (Chen and Rowe 1999).

LifePaths is intended to incorporate socio-economic information from all relevant sources available to Statistics Canada. Consequently the construction of the model has motivated research into application of methodologies for exploiting multiple data sources. Embedding an estimated model in LifePaths is a powerful tool for deriving implications of the model that can be compared to information

¹ Ian Cahill, Partnership and Continuous Evaluations, HRDC, 140 Promenade du Portage, Phase IV 3rd floor, Room 3D475, Gatineau, Québec K1A 0J9, and Edward J. Chen, Household Survey Methods Division, Statistics Canada, R.H. Coats Building 16th floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

from other sources. For example, Rowe and Lin (1999) derived job tenures by simulation from a model estimated using short-period longitudinal data, then compared the results with data from a cross-sectional survey. We report on one aspect of the continuing effort to build a tool providing the maximum information that can be extracted from Statistics Canada's data sources.

The paper is organized to illustrate the way in which technical problems are often encountered in the course of building LifePaths, and how their solution is integrated with the model development process. To do this, a fair amount of background detail on associated issues is provided. Section 2 outlines the context of the benchmarking problem, and section 3 presents the theory behind our solution, with some possible extensions for further work. Section 4 describes the models to which it will be applied, including some details concerning the estimation of their parameters in the base period, then section 5 describes the application of the benchmarking method to these models. We display and discuss our empirical results in section 6, then close with some overall conclusions in section 7.

2. CONTEXT OF THE PROBLEM

We provide context in this section by presenting an overview of the LifePaths model structure, a brief description of data sources involved, and a discussion of how the benchmarking problem arose.

2.1 Structure of the LifePaths Model

The LifePaths model simulates individual lifetimes as a series of events which modify the set of "state variables" describing the demographic, social, and economic circumstances of the individual. Waiting times to every possible event are associated with an individual, although they may be infinite. The waiting times may be conditioned on the values of state variables. The event type with the shortest waiting time occurs (its associated functions are called). Modification of any state variable at the occurrence of an event may lead to the generation of new waiting times for other events.

LifePaths initialises a case by randomly generating a "dominant" individual's sex, province of residence, age at immigration and year of birth. The year of birth can range from 1892 to 2051. Mortality and immigration assumptions are designed to reproduce provincial age-sex structures. When a dominant individual marries, enters a common-law union, or has a child, a non-dominant individual of suitable characteristics is created and is linked to the dominant individual, forming part of the case. Once created, non-dominant individuals undergo the same possible events as dominant individuals. However, since their purpose is to complete the profile of the dominant actor, they are usually filtered from all tabular reports.

LifePaths presently includes models of fertility, mortality, marriage (including common-law unions), educational careers, labour force careers, maternity leave, hours of work, earnings, taxes, and transfers. The model of the labour force careers describes transitions between the states "paid employee," "self-employed," and "not employed." It also includes a model of retirement and student work. The model of secondary and post-secondary educational careers at the provincial level is mature and highly developed.

2.2 The Data Sources

The estimation of base parameters for the model of maternity leave was carried out using data from SLID covering maternity leaves beginning in the period 1993-1996. Using data from 1997 allowed us to follow most maternity leaves to completion rather than using extensively censored data. This is a household survey designed to permit both longitudinal and cross-sectional analysis of people's financial and work situations. Starting in 1993, SLID follows the same respondents for six years, with new rotation groups introduced every three years. Each rotation group includes about 15,000 households with 30,000 adults. From this survey we obtain the month of child birth, monthly data on labour force status, and a rich set of explanatory variables including job tenure, an indicator of self-employment, birth order of the child, presence of an employed spouse, province of residence, education level, and age. We can also determine if a mother who left a job within 4 months of birth has returned to the same job within 16 months. This is used as a practical definition of maternity leave and becomes our unit of analysis, with a slight expansion to include the 1% of cases where a mother returned to a different job from a labour market state of absence in the previous month. Using this unit of analysis we get a sample size of 835 births. As we show in section 6, this sample size is adequate to reveal some key explanatory factors. More precisely, several factors are found to be significant at the 95% confidence level. This sample contains about 730 unique mothers, representing over 87% of the sample of births. This means that there will be some correlation between observations as a result of those mothers who have two or more maternity leaves within the observation period, but we did not feel that it is of sufficient magnitude to warrant any special statistical tools.

The LFS is a monthly household survey focussing on labour force status, and also reporting a number of demographic characteristics. The survey is normally used exclusively for cross-sectional analysis. For the LifePaths project, however, a file covering the period from 1976 to 1995 was constructed that follows individuals as they rotate through the six monthly rotation groups of the survey, providing a six-month window on each individual's labour market activity. Since the number and ages of children are recorded each month, it is possible to observe the

appearance of a new child. Since all surveys throughout the period are used, the sample size is very large, and about 26,000 births are observed.

In the LFS window we note the labour force status of a new mother when the child is first reported. This is the key to estimating the probability of choosing a maternity leave, rather than leaving the labour force. We begin by considering $P(E)$, the proportion of such mothers who are employed. If the mother is "employed, at work," we suppose that they took a brief absence from their job – less than a month. If they are "employed, absent from work," it may be that they have chosen to take a maternity leave absence from their job and then return to it. However this may not always be the case. A new mother who we observe as employed and absent (EA) may later make a transition out of employment (to NE). To correct for this, considering mothers with a child of age less than a year observed in a window, we calculate the proportion $P(EA-NE)$ of transitions out of the "employed, absent from work" state that are to a not-employed state. We also estimate the proportion $P(NE-OJ)$ of mothers who return to an old job (OJ) after having left employment. The estimate is obtained by using observations on mothers with a young child who make transitions from a not-employed state to a job with a start date earlier than the previous month. Our estimate of the probability of choosing a maternity leave is now $P(E) - P(EA - NE) + P(NE - OJ)$.

It is also possible to observe mothers with a child of age less than a year making a transition from the status "employed, absent from work for personal or family responsibilities" to the status "employed, at work." We use this transition as a proxy for the return to work after a maternity leave. Since the duration of absence is reported in the previous month, this is the key to benchmarking the survival model.

The preceding discussion illustrates the weakness of the LFS data for a study of maternity leave, relative to SLID data. In addition to having fewer explanatory variables available than in SLID, we must accept proxies for the dependent variables. Nevertheless, we require the historical depth of the LFS. This relationship between the data sets is the context of the benchmarking problem described in the next section.

Both the SLID and the LFS have complex sample designs involving detailed stratification, and complex methods for calculating observation weights. We always make use of observation weights, both in estimation and in the calculation of frequencies. The methods used are fairly simple, and are discussed in sections 4 and 5.

2.3 The Benchmarking Problem

The context of our benchmarking problem is a model of women choosing between leaving the labour force or taking a maternity leave, and if they choose a leave, deciding how long that leave should be. The first decision is represented by a binary logit model, and the second by a semiparametric

survival model, both including a vector of explanatory variables and associated parameters. In LifePaths, the decisions are made as part of the maternity leave choices event, which always occurs in the middle of a pregnancy. SLID is quite adequate for estimation of the base parameters of both these models. However, since a major goal of the LifePaths project is to incorporate historical patterns of change in socio-economic processes, it was necessary to benchmark the SLID parameter estimates to annual estimates of dependent variable means obtained from the LFS.

In this problem, we assume stable observed characteristics of the population. There are two reasons for this. First, LifePaths is a work in progress, and the benchmarking exercise we report on was carried out at a stage when other parts of the model that predict these characteristics were being extensively revised. In section 3.3, we touch on the consequences of evolving population characteristics. Second, we suppose that the primary reason for systematic change in observed outcomes between time periods is change in some factors not included in the measured characteristics of individuals. In the case of our application we observed a trend towards choice of maternity leave over leaving the labour force which seems to be due to social change rather than changes in the composition the population. We also observed a change in the distribution of maternity leave durations that appears to be due to changes in the Unemployment Insurance (UI) program implemented in Bill C-21 in 1990. At that time Parental Benefits were introduced, which extended the period during which many mothers could receive benefits from 15 to 25 weeks. Many mothers return to work at a time close to when they have exhausted UI benefits.

3. BENCHMARKING METHODOLOGY

In this section we present the method in an abstract form in order to clarify the assumptions, develop notation, and to reveal the similarity between the application to binary choice and to survival analysis.

3.1 Application to Binary Choice

The basic model for the benchmarking methodology relates to binary choice. Since we are not primarily interested in changes in the population, we simplify the analysis by assuming that the explanatory variables or individual characteristics in period τ are represented by a series of independent identically distributed random vectors X^τ . We recognise that this is quite a strong assumption. Nevertheless, for the reasons discussed in section 2.3, we use it as our empirical work. Section 3.3 shows that it is a fairly simple matter to extend the theory to incorporate trends in the independent variables.

Consider a linear predictor given by

$$\eta^\tau(x) = \beta'x + \gamma^\tau \quad (3.1)$$

where β is a vector of coefficients constant over time, x is a possible outcome of X^τ , and γ^τ represents a parameter specific to period τ . Notice that x contains no “constant term.” Let Y^τ be a random variable, jointly distributed with X^τ , that takes the values 1 if an event occurs and 0 if it does not. Suppose that the probability of the event, conditional on characteristics x , is given by

$$E(Y^\tau | X^\tau = x) = \pi^\tau(x) = F(\eta^\tau(x)) \quad (3.2)$$

where we require F to be a continuous distribution function. The values of the function will then be bounded by zero and one, and it will have an inverse g , so that

$$\eta^\tau(x) = g(\pi^\tau(x)). \quad (3.3)$$

In the context of generalised linear models, g is called a link function. We begin by finding maximum likelihood estimates of the base parameters β and $\hat{\gamma}^{\tau_0}$ using data for the time period τ_0 (in our case this is the period when SLID data are available). Of course these data must include variables corresponding to outcomes of both X^τ and Y^τ . It remains to estimate γ^τ for each period τ . Equations (3.1) and (3.3) imply that

$$E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} = \gamma^\tau - \gamma^{\tau_0} = E\{g(\pi^\tau(X^\tau)) - E\{g(\pi^{\tau_0}(X^{\tau_0}))\}\}. \quad (3.4)$$

Since we have observations only on the outcomes of Y^τ from the LFS for every period, we estimate the terms γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) \quad (3.5)$$

where $\hat{\pi}^\tau$ is an estimate of $E(Y^\tau)$. Using the LFS, this estimate is the weighted frequency of the event in the time period τ (taking each weight from the month where a child is first observed). To justify this procedure we use equation (3.4) and assume an approximation

$$E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \approx g(E\{\pi^\tau(X^\tau)\}) - g(E\{\pi^{\tau_0}(X^{\tau_0})\}). \quad (3.6)$$

Inaccuracy will arise due to Jensen’s inequality in regions where g is convex or concave. Nevertheless, if g can be locally approximated by a linear function in the regions where $\pi^\tau(X^\tau)$ and $\pi^{\tau_0}(X^{\tau_0})$ are concentrated, then (3.6) may be quite accurate. The fact that g has an inflection point at 0.5 may aid the approximation when probabilities are dispersed around this value.

Fortunately we are able to test the adequacy of the estimator by simulating the estimated model in LifePaths and comparing the predicted frequencies of the event with corresponding weighted frequencies observed in the data. The results indicate that it is quite adequate for our application.

3.2 Application to Survival Analysis

We will show in section 5.2 that the approach outlined above can also be extended for use with a semiparametric survival model by adding an index t representing the duration in the current state, so that (3.5) becomes

$$\hat{\gamma}^\tau(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) \quad (3.7)$$

where $\hat{\pi}^\tau(t)$ represents the empirical hazard function.

3.3 Trends in the Independent Variables

The benchmarking method may be improved by taking the changes in observed characteristics into account. As we noted in section 2.3, this would be considered when other parts of LifePaths are in a more mature form. To do this we relax the assumption that the random vectors X^τ are identically distributed. Equation (3.4) then becomes

$$\begin{aligned} E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} &= \gamma^\tau - \gamma^{\tau_0} + \beta' \{E(X^\tau) - E(X^{\tau_0})\} \\ &= E\{g(\pi^\tau(X^\tau)) - E\{g(\pi^{\tau_0}(X^{\tau_0}))\}\} \end{aligned} \quad (3.8)$$

Based on this, we might estimate γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) - \hat{\beta}'(\bar{x}^\tau - \bar{x}^{\tau_0}) \quad (3.9)$$

where \bar{x}^τ is the vector of mean values of the characteristics in period τ . Of course it may not be possible to obtain all of the mean values from the same data source. The method would extend to the survival model case in the same manner as (3.7) to give

$$\begin{aligned} \hat{\gamma}^\tau(t) &= \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) \\ &\quad - \hat{\beta}'(\bar{x}^\tau(t) - \bar{x}^{\tau_0}(t)). \end{aligned} \quad (3.10)$$

4. MODELS AND THE ESTIMATION OF BASE PARAMETERS

As explained in section 3.1, the base parameters β and $\hat{\gamma}^{\tau_0}$ are estimated by maximum likelihood using data from the period τ_0 . We use data from SLID on all maternity leaves beginning in the period 1993–1996 (our base period τ_0). We do not attempt to estimate annual changes in the constant term γ throughout this period.

4.1 The Binary Logit Model

We adopt the logit model to represent a mother’s choice between taking a maternity leave and withdrawing from the labour force. From now on we adopt a more conventional econometrics notation and use a subscript i to index a random variable or outcome associated with an individual i . We suppose that a random variable Y_i^τ takes values 0 or 1, with $Y_i^\tau = 1$ indicating that new mother i with vector of characteristics x_i in period τ chooses to take a maternity leave, conditional on her having been employed, and that

$$\pi_i^\tau = P(Y_i^\tau = 1) = F(\eta_i^\tau) = \frac{\exp(\eta_i^\tau)}{1 + \exp(\eta_i^\tau)} \quad (4.1)$$

where $\eta_i^\tau = \beta' x_i + \gamma^\tau$ is the linear predictor of equation (3.1) and F is the logistic distribution function. We estimate the base parameters β and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\beta, \gamma^{\tau_0})$ where

$$\begin{aligned} L(\beta, \gamma^\tau) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N) \\ &= \prod_{y_i=0} [1 - F(\eta_i^\tau)] \prod_{y_i=1} F(\eta_i^\tau) \\ &= \prod_i [F(\eta_i^\tau)]^{y_i} [1 - F(\eta_i^\tau)]^{1-y_i} \end{aligned} \quad (4.2)$$

and

$$\begin{aligned} \ln L(\beta, \gamma^\tau) &= \sum_i \{ y_i \ln F(\eta_i^\tau) \\ &\quad + (1 - y_i) \ln [1 - F(\eta_i^\tau)] \}. \end{aligned} \quad (4.3)$$

Longitudinal SLID weights in the year of the child's birth are scaled to sum to the sample size, and are then used to weight the terms of the log-likelihood and its derivatives. The weighted score equations are

$$\begin{aligned} \frac{\partial L(\beta, \gamma^\tau)}{\partial \beta} &= \sum_i w_i x_i y_i - \sum_i w_i x_i F(\eta_i^\tau) = 0 \\ \frac{\partial L(\beta, \gamma^\tau)}{\partial \gamma^\tau} &= \sum_i w_i y_i - \sum_i w_i F(\eta_i^\tau) = 0. \end{aligned} \quad (4.4)$$

The solution, which maximises the log-likelihood, was found by Newton-Raphson iteration. The logit model has been used often by statisticians and econometricians, and there is an extensive literature. For example, see Chambless and Boyle (1985), Roberts, Rao, and Kumar (1987), and Morel (1989).

4.2 The Semiparametric Survival Model: Basic Form

For mothers who have chosen to take a maternity leave from their job, we use a survival model to describe the duration of their leave. The probability density function (pdf) of the distribution has a complex shape, as can be seen from the graphs in section 6.4. There is spike at durations of less than a month and a mode which appears to represent the maximum Unemployment Insurance special benefits entitlement available to mothers after 1990 (15 weeks of Maternity Benefits, plus 10 weeks of Parental Benefits, plus a two-week waiting period). We began the study by estimating various fully parametric models, including a log-logistic survival model combined with a logit model to

predict durations of less than a month, but were unable to obtain an adequate fit. To solve this problem, we follow Prentice and Gloeckler (1978), Han and Hausman (1986) and Meyer (1990), by nonparametrically estimating the effect of time on the hazard of returning to work. The hazard of returning to work is specified in a proportional hazards form:

$$\lambda_i^\tau(t) = \lambda_0^\tau(t) \exp \{ \beta' x_i(t) \} \quad (4.5)$$

where $\lambda_0^\tau(t)$ is the unknown baseline hazard at leave duration t and time period τ , $x_i(t)$ is a vector of explanatory variables for mother i , and β is a vector of coefficients. The data tell us which of the intervals $[0,1)$, $[1,2)$, $[2,3)$, ... contains the spell duration (in our case the units are months), and the model can be interpreted as an incompletely observed continuous time hazard model with no restriction on the form of the baseline hazard. If T_i^τ is the duration of leave for mother i during period τ , then for $t = 1, 2, 3, \dots$, the probability that the spell lasts until time t , given that it has lasted until $t-1$, can be written as

$$\begin{aligned} P(T_i^\tau > t | T_i^\tau \geq t-1) &= \exp \left[- \int_{t-1}^t \lambda_i^\tau(u) du \right] \\ &= \exp \left[- \exp \{ \beta' x_i(t) \} \int_{t-1}^t \lambda_0^\tau(u) du \right] \end{aligned} \quad (4.6)$$

if we assume that $x_i(t)$ is constant on the interval between $t-1$ and t . In order to apply the theory of section 3, we can rewrite equation (4.6) as

$$\begin{aligned} 1 - \pi_i^\tau(t) &= P(T_i^\tau \geq t | T_i^\tau \geq t-1) \\ &= \exp \{ - \exp \{ \beta' x_i(t) + \gamma^\tau(t) \} \} \\ &= \exp \{ - \exp \{ \eta_i^\tau(t) \} \} \end{aligned} \quad (4.7)$$

where

$$\gamma^\tau(t) = \ln \left[\int_{t-1}^t \lambda_0^\tau(u) du \right]. \quad (4.8)$$

One may censor any ongoing observations at some large duration T . Again we can estimate the base parameters β and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\gamma^{\tau_0}, \beta)$. Since we will always be referring to data from the base period for the remainder of section 4, we drop superscripts τ_0 .

The likelihood function is given by

$$\begin{aligned} L(\gamma, \beta) &= \prod_{i=1}^N [[1 - \exp \{ - \exp(\eta_i(k_i)) \}]^{\delta_i} \\ &\quad \prod_{t=1}^{k_i} \exp \{ - \exp(\eta_i(t)) \}] \end{aligned} \quad (4.9)$$

where $\gamma = [\gamma(1), \gamma(2), \dots, \gamma(T)]'$, C_i is a censoring time, $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise, $k_i = \min(\text{int}(T_i), C_i)$. The log-likelihood is therefore

$$\ln L(\gamma, \beta) = \sum_{i=1}^N [\delta_i \ln\{1 - \exp\{-\exp(\eta_i(k_i))\}\} - \sum_{t=1}^{k_i} \exp(\eta_i(t))]. \quad (4.10)$$

Weights from the months that a child is first observed are scaled to sum to the sample size, and then used to weight the terms of the log-likelihood function and its derivatives. The weighted log-likelihood function is maximised by the quasi-Newton algorithm of Broyden, Fletcher, Goldfarb, and Shanno (BFGS), using an implementation based on Dennis and Schnabel (1983).

4.3 The Semiparametric Survival Model: with Work-to-Birth Gap Decision

The situation in our application is complicated somewhat by our desire to model the duration from leaving the job until the birth (the work-to-birth gap), as well as the hazard of returning to work from a maternity leave. The model of work-to-birth gap is estimated separately, based on SLID data. Examination of the mean gap duration for each year in the LFS data indicates that this duration has been fairly stable over time, so the model is not benchmarked. Nevertheless, a modification of the semiparametric survival model is necessary to incorporate the separate model of work-to-birth gap. This can be accomplished by assuming that the work-to-birth gap decision, possibly involving health considerations, acts to constrain the desired total duration. This means that the above model would apply to the desired total duration, which is unobservable, and might be labelled T^* .

In cases where the desired duration was shorter than the work-to-birth gap, the mother might return to work as soon as possible after the birth. This means that in cases where we observe a significant work-to-birth gap (greater than a month), and the mother returns soon after birth (within a month), all that is known about desired duration is that

$$T^* \leq T$$

where T is the total duration of leave. This is equivalent to a situation labelled “left censoring” by Cox and Oaks (1984, page 178), where observation does not start immediately and some individuals have already failed before it does.

From such an observation we get a contribution to the likelihood function and its logarithm given by

$$L_i = 1 - \prod_{t=1}^{k_i} P(T^* \geq t | T^* \geq t-1) = 1 - \prod_{t=1}^{k_i} \exp[-\exp(\eta_i(t))] \quad (4.11)$$

and

$$\ln(L_i) = \ln\{1 - \exp[-\sum_{t=1}^{k_i} \exp(\eta_i(t))]\}. \quad (4.12)$$

Unfortunately the log-likelihood expression does not simplify like the corresponding expression for “right-censored” observations. In spite of this, Monte Carlo experiments indicate that estimation is not a problem even in heavily censored data sets.

Longitudinal SLID weights in year of the child’s birth are used in same manner as for the basic form of the survival model.

5. BENCHMARKING THE MODELS

To begin the benchmarking procedure we must invert the distribution function F given in equation (3.2) to find the link function g . We then apply equation (3.5) in the case of the logit model, and equation (3.7) in the case of the survival model.

5.1 Application to the Binary Logit Model

To benchmark the logit model we first invert the logistic distribution function in equation (4.1) to obtain

$$\eta_i^\tau = g(\pi_i^\tau) = \ln\left(\frac{\pi_i^\tau}{1 - \pi_i^\tau}\right) \quad (5.1)$$

where g is the well-known logit function. We can then apply equation (3.5) and (5.1) to obtain

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) = \hat{\gamma}^{\tau_0} + \ln\left(\frac{\hat{\pi}^\tau/(1 - \hat{\pi}^\tau)}{\hat{\pi}^{\tau_0}/(1 - \hat{\pi}^{\tau_0})}\right) \quad (5.2)$$

where for $\tau < \tau_0$, each $\hat{\pi}^\tau$ is the frequency of choosing maternity leave calculated from LFS data for maternity leaves beginning in year τ , and $\hat{\pi}^{\tau_0}$ is the frequency from SLID data.

5.2 Extension to the Survival Model

From equation (4.7) we get

$$\pi_i^\tau(t) = 1 - \exp[-\exp\{\eta_i^\tau(t)\}] = F\{\eta_i^\tau(t)\} \quad (5.3)$$

where

$$\eta_i^\tau(t) = \beta' x_i(t) + \gamma^\tau(t). \quad (5.4)$$

In this case F is an extreme value distribution that is easily inverted to obtain

$$\eta_i^\tau(t) = \ln[-\ln(1 - \pi_i^\tau(t))] = g(\pi_i^\tau(t)). \quad (5.5)$$

For benchmarking we can use equation (3.7) with the observed frequencies in period τ represented by the empirical hazard or occurrence/exposure ratio given by

$$\hat{\pi}^{\tau}(t) = d^{\tau}(t) / r^{\tau}(t) \tag{5.6}$$

where, for spells beginning in period τ , $d^{\tau}(t)$ is the number of mothers who fail in the interval $(t - 1, t]$ and $r^{\tau}(t)$ is the number of mothers in view at duration t , including those censored at time t (censoring can only occur at the end of intervals). Numbers of mothers were calculated from sample counts by applying the LFS weight from the month that a new mother returns to work. The empirical hazard and the corresponding estimator for the survivor function implied by the product law of probabilities were studied by Kaplan and Meier (1958). The use of the empirical hazard in equation (3.7) together with equation (5.5) yields

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + \ln \left(\frac{\ln[1 - \hat{\pi}^{\tau}(t)]}{\ln[1 - \hat{\pi}^{\tau_0}(t)]} \right). \tag{5.7}$$

6. EMPIRICAL RESULTS

The results of estimation in the base period, and the results of simulation with benchmarked parameter estimates are presented for both models. The simulation results are compared with annual survey sample frequencies of choosing a maternity leave in the case of the logit model, and with annual survey frequency distributions of maternity leave duration in the case of the survival model.

6.1 Estimation Results for the Binary Logit Model

The estimation results obtained from estimating the logit model from SLID data are presented in Table 1. Omitted dummy variable categories, which form the reference categories for the variables used in the model, were province of residence Ontario and highest education level “some post secondary.” Individual and family income variables were tested, but were found not to be significant, and so were not included in the regression.

There may be some bias in the estimates, particularly those of the standard errors, due to the fact that the complex SLID sample design was accounted for only through the weights applied to the log-likelihood.

The significant positive effect of job tenure seems reasonable for a number of reasons. A lengthy tenure might indicate that the woman has acquired firm-specific human capital and has achieved some seniority. It would also be an indicator of strong attachment to the labour force generally. On the firm side, the longer the woman’s job tenure, the longer the leave that the firm is likely to grant with a guarantee that she can return to her job. Also, provincial government guarantees of job security also depend on job tenure. Finally, a lengthy job tenure means that the woman will likely meet the Unemployment Insurance eligibility requirements (20 weeks of insured employment). A dummy variable indicating that UI entrance requirements were met was tested and found to be just significant at the 5% level. However, because we are not able at this stage to model

changes in the UI program through the influence of covariates, because of uncertainty in interpretation, and because of high correlation with job tenure, it was not included. In the LFS, self-employed workers are reported as having a transition out of employment only when they terminate their business. Since taking a leave simply means not terminating the business, a significant positive effect for the indicator of self-employment is to be expected. Having been self-employed before the birth increases the odds of taking a maternity leave by 333%, the strongest effect that we see for an indicator variable.

Table 1
Binary Logit Parameter Estimation Results

Parameter	Estimate of Coefficient	Contribution to Odds Ratio*	Std Error of Coefficient	Prob-Value
Constant	-6.432	0.002	2.995	0.0318
NFLD	-0.829	0.436	0.741	0.2636
PEI	0.931	2.537	1.612	0.5633
NS	-0.456	0.634	0.541	0.3992
NB	0.207	1.230	0.675	0.7596
QUE	-0.361	0.697	0.247	0.1437
MAN	-0.490	0.613	0.503	0.3306
SASK	-0.163	0.850	0.458	0.7218
ALTA	-0.200	0.819	0.325	0.5379
BC	-0.120	0.887	0.300	0.6899
Job Tenure (mths)/10	0.094	1.099	0.026	0.0003
Self-employed?	1.203	3.330	0.418	0.0040
Age (Years)	0.479	1.614	0.199	0.0160
(Age^2)/10	-0.071	0.931	0.033	0.0296
< High School Grad	-0.702	0.496	0.357	0.0490
High School Grad	-0.148	0.862	0.276	0.5913
University Grad	-0.292	0.747	0.229	0.2027
First Child?	-0.525	0.592	0.192	0.0063

log-likelihood = -381.553
Number of Observations = 835
Observations are given the SLID longitudinal weight from the year of birth, scaled to sum to the sample size

* This is the exponential of the coefficient. It may be interpreted as the proportional change in the odds ratio due to a unit change in the corresponding independent variable.

The effect of the first child indicator also seems reasonable. The odds for maternity leave for a first-time mother is only 59% of the odds for maternity leave for a mother of more than one child, given that all other characteristics are the same – i.e. first-time mothers are more inclined to job separation than the mothers who already have children. This may be partly a consequence of the fact that our sample consists of mothers who have been employed within 4 months of the birth. Mothers who have more than one child tend to space them within a few years at most. If they are employed just before a second or subsequent births, they will have already demonstrated that they returned to work after an absence that must have been less than the gap between births. This at least rules out some common patterns of withdraw from the labour force – for example staying at home until all children are in school.

The effect of age is more difficult to interpret since the effect on the log-odds ratio is non-linear. By drawing a graph of the term $-.479 * age - .0071 * age^2$ one can see that, as age increases, the log-odds of taking a maternity leave first increases, but that the rate of increase declines until a level point at the maximum log-odds is reached by the age of 34. Since the number of mothers declines considerably after this age, the subsequent decline may not be meaningful. One might hazard a conjecture that, among young mothers, being relatively older indicates more attachment to the labour force and thus a stronger tendency to take a maternity leave, while among older mothers, who are past the stage of first entering the labour force, this effect is reduced. However, the results are probably not precise enough to draw any firm conclusion about this.

6.2 Simulation Results for the Benchmarked Binary Logit Model

The benchmarking exercise consists of adjusting the constant term of the model in the manner described by (5.2) for each year in the period 1975-1992. The constant term is not adjusted after 1992, partly because the LFS data do not indicate a strong trend after 1992. The model is then incorporated in LifePaths and a simulation is run. For each year from 1976 to 1995, Figure 1 shows both the frequency of choosing a leave in the LifePaths simulation, and the frequency estimated from the LFS. For the period 1993-1995, estimates from SLID are also presented.

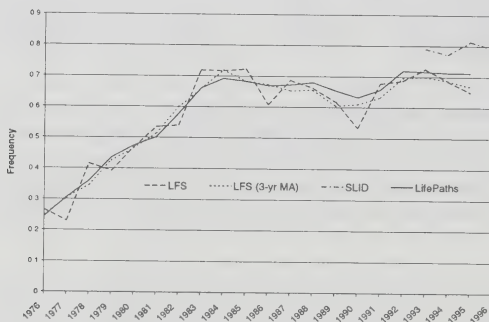


Figure 1. Frequency of Choosing a Maternity Leave 1976-1996

The simulation captures the change over time revealed by the LFS data during the period 1976-1992. There is no benchmark adjustment implemented in the LifePaths simulation after 1992, so that the base parameters estimated from pooled SLID data 1993-1996 are effective. The simulated frequency is slightly lower than the observed SLID frequency during this period. Two possible sources of error are an insufficiently flexible specification of the binary choice model, and differences between the SLID estimates of explanatory variables and those provided by LifePaths.

6.3 Estimation Results for the Survival Model

The results obtained from estimating the semiparametric survival model from SLID data are presented in Table 2. As in the binary logit model estimation, omitted dummy variable categories were province of residence Ontario and highest education level "some post secondary." Since the dependent variable is the hazard of returning to work, a positive coefficient for a covariate indicates an influence that tends to shorten the duration of maternity leave.

The estimates of the constant terms in the duration-dependent linear predictor given by (4.7) are denoted in Table 2 by GAMMA_i , $i = 1, 2, \dots, 15$. This represents the influence of the baseline hazard incorporating the influence of duration.

Table 2
Survival Model Parameter Estimation Results

Parameter	Estimate	Std Error	Prob-Value
Job Tenure (mths) /10	-0.030	0.010	0.0024
NFLD	0.195	0.426	0.6470
PEI	0.307	0.490	0.5313
NS	0.173	0.253	0.4940
NB	0.109	0.293	0.7091
QUE	0.111	0.117	0.3411
MAN	-0.402	0.253	0.1116
SASK	-0.303	0.213	0.1539
ALTA	0.270	0.154	0.0798
BC	-0.440	0.148	0.0030
Self-Employed?	1.665	0.157	0.0000
Age	-0.253	0.041	0.0000
Age** 2 / 10	0.043	0.007	0.0000
First Child?	-0.301	0.090	0.0009
< High School Grad	0.508	0.206	0.0135
High School Grad	-0.124	0.125	0.3212
University Grad	-0.374	0.108	0.0006
Employed Spouse?	0.109	0.151	0.4703
Gamma1	2.570	0.609	0.0000
Gamma2	-1.136	0.816	0.1636
Gamma3	-0.466	0.719	0.5176
Gamma4	0.780	0.640	0.2232
Gamma5	1.425	0.627	0.0231
Gamma6	2.755	0.613	0.0000
Gamma7	3.640	0.612	0.0000
Gamma8	3.413	0.620	0.0000
Gamma9	3.465	0.630	0.0000
Gamma10	3.387	0.649	0.0000
Gamma11	4.579	0.655	0.0000
Gamma12	4.285	0.785	0.0000
Gamma13	3.645	1.110	0.0010
Gamma14	3.746	1.281	0.0034
Gamma15	6.215	2.415	0.0101

log-likelihood = -1165.06

Number of Observations 3411

Observations are given the SLID longitudinal weight from the year of birth, scale to sum to the sample size

Again, individual and family income variables were tested and found not to be significant. Both this finding and the importance of a self-employment indicator as a predictor of early return to work accord with the findings of Marshall (1999). Marshall found that education variables were not significant in determining whether a mother would return to work within a month. We find however, that university graduation has a significant negative effect on the hazard (positive effect on duration). Job tenure has a significant negative effect on the hazard, possibly reflecting its relationship with Unemployment Insurance entitlement and job security.

6.4 Simulation Results for the Benchmarked Survival Model

In the case of the semiparametric survival model, benchmarking consists of adjusting all of the terms $GAMMA_i$, $i = 1, 2, \dots, 15$ of the previous section according to (5.8) for each of the years in the period 1975-1992. The model is then simulated as part of LifePaths.

The frequency distribution of simulated maternity leave durations is presented and compared to the corresponding observed frequency distribution from LFS data. In order to present the results, the frequencies in 3-year periods were averaged. A key feature of the frequency distribution is an abrupt change apparently due to the introduction of parental benefits with Bill C-21 at the end of 1990. Since mothers with maternity claims in progress at the time of implementation were entitled to parental benefits, the claims beginning in 1990 represent a mixture of regimes. For this reason the year 1990 is not included in any of the 3-year averages. In Figures 2 and 3 we use disjoint 3-year periods covering 1976-1984. To balance periods before and after 1990 using available data, in Figures 4 and 5 we use the overlapping periods 1985-1987, 1987-1989, 1991-1993, and 1993-1995.

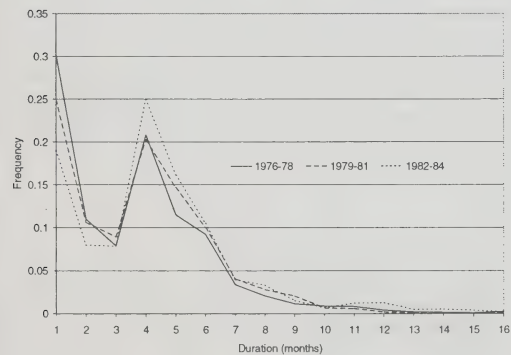


Figure 2. LifePaths: Distribution of Leave Durations for 1976-1984

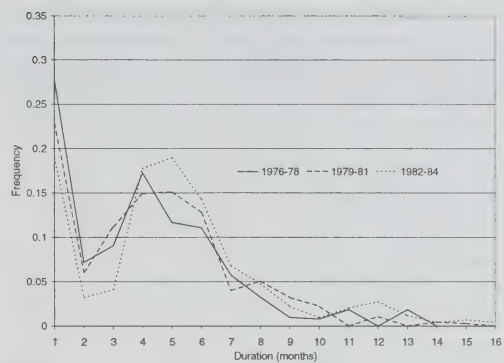


Figure 3. LFS Data: Distribution of Leave Durations for 1976-1984

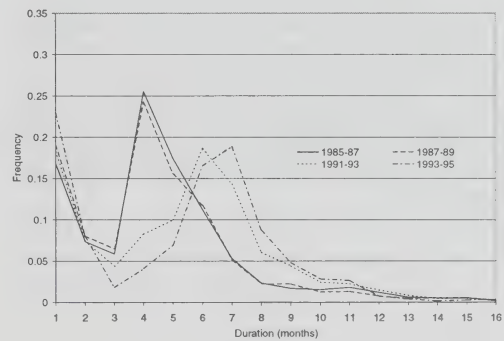


Figure 4. LifePaths: Distribution of Leave Durations for 1985-1989 and 1991-1995

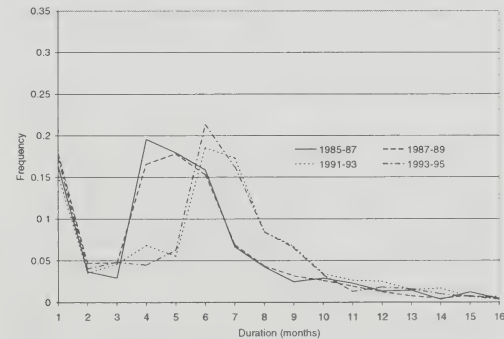


Figure 5. LFS Data: Distribution of Leave Durations for 1985-1989 and 1991-1995

The distribution of durations derived from SLID data 1993-1996 is presented in Figure 6. This may be compared with the simulated data shown in Figure 4 for the period 1993-1995, since no benchmarking is applied after 1992.

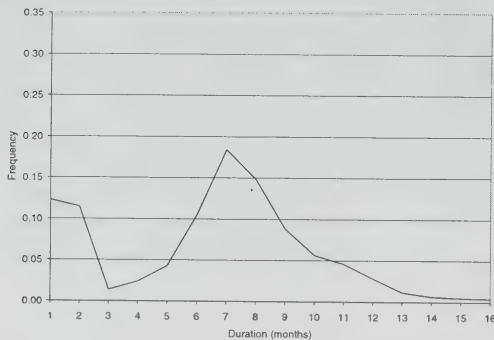


Figure 6. SLID Data: Distribution of Leave Durations for 1993-1996

In Figure 7 we present the average duration of maternity leaves beginning in each year of the observed period. The average of simulated durations are compared with those from the surveys.

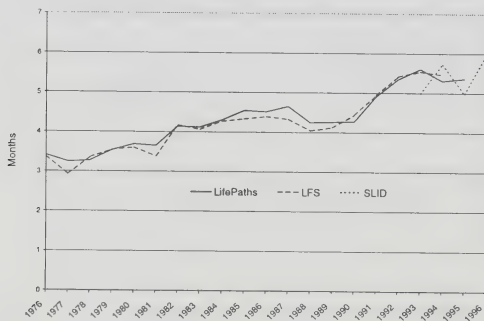


Figure 7. Average Duration of Maternity Leave 1976-1996

6.5 Evaluation of Benchmarking Performance

The benchmarking method appears to be very effective in the case of the binary logit model. The trend of the LFS data is well reflected in the LifePaths simulation. In the case of the survival model, the key feature of the LFS data is the abrupt shift of the mode of the frequency distribution after 1990, apparently due to the introduction of parental benefits. This shift has been captured by the simulated data. Also the average duration of maternity leave in the simulation fits the LFS data very closely.

A noticeable divergence between the simulation and the LFS data is the height of the mode at the interval (3, 4] months in the frequency distribution of the durations from

LifePaths from 1982-1989. This may be due to the effect of trends in the values of explanatory variables, which we have assumed to be stable. Further work is necessary to establish this. A possible extension to the model was discussed in section 3.3.

7. CONCLUSIONS

The technique that we have developed appears to be quite successful in benchmarking of the logit and survival model parameters so that the essential features of the LFS data are captured in LifePaths predictions. The key to benchmarking the logit model is the adjustment of the parameter corresponding to the "constant term" in the linear predictor that is imbedded in the logistic distribution function in order to predict the conditional expectation of the dependent variable. Section 3.1 develops the technique in a general framework that includes other models of binary choice. Particularly, it would extend to the popular probit model where a linear predictor is embedded in the standard normal distribution function. Benchmarking of the semiparametric survival model hinges on the adjustment of all the parameters representing the baseline hazard. Our results illustrate how the entire shape of the distribution of durations predicted by the model can be made to evolve through time according to a pattern revealed by supplementary data.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to Steve Gribble and members of the Socio-economic Modelling Group at Statistics Canada for useful comments throughout the development of the maternity leave module, to Geoff Rowe and Huan Nguyen for use of their computer program to follow individuals through rotations in the LFS, to Katherine Marshall for advice on the use of SLID and for sharing computer programs, to Adrienne ten Cate for fruitful discussions, and to an anonymous referee for several improvements. This work was performed when both authors worked in the Socio-Economic Modeling Group, Statistics Canada, R.H. Coats Building 24th floor, Tunney's Pasture

REFERENCES

- APPLEBY, J., BOOTHBY, D., ROULEAU, M. and ROWE, G. (1999). Level and Distribution of Individual Returns to Post-Secondary Education: Simulation Results from the LifePaths Model. Presented at the 1999 meetings of the Canadian Economics Association.
- CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum Likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, A: Theory and Methods*, 14, 177-192.

- CHEN, E.J., and ODERKIRK, J. (1997). Varied Pathways: The Undergraduate Experience in Ontario, Feature article. *Education Quarterly Review*, Statistics Canada, 4, 3, 47-62.
- CHEN, E.J., and ROWE, G. (1999). Trend Correlation of Labour Market Earnings in Canada: 1982 to 1995. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 173-179.
- COX, D.R., and OAKS, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.
- DENNIS, J.E. Jr, and SCHNABEL, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.
- HAN, A., and HAUSMAN, J. A. (1986). Semiparametric Estimation of Duration and Competing Risk Models. M.I.T. Working Paper No. 450.
- KAPLAN, E.L., and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-81.
- MARSHALL, K. (1999). Employment after childbirth. *Perspectives on labour and income*. Statistics Canada, Autumn 1999, 18-25.
- MEYER, B.D. (1990). Unemployment Insurance and Unemployment Spells. *Econometrica*, 58, 757-782.
- MOREL, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology* 15, 205-223.
- PLAGER, L., and CHEN, E.J. (1999). Student Debt from 1990-91 to 1995-96: An Analysis of Canada Student Loans Data. MAJOR RELEASES, *THE DAILY* and *Education Quarterly Review*, Statistics Canada, 5, 4, 10-35.
- PRENTICE, R., and GLOECKLER, L. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data, *Biometrics*, 34, 57-67.
- ROBERTS, G.A., RAO, J.N.K. and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROWE, G., and CHEN, E.J. (1998). An Increment-Decrement Model of Secondary School Progression for Canadian Provinces. *Proceedings: Symposium on Longitudinal Analysis for Complex Surveys*, Statistics Canada, 167-178.
- ROWE, G., and LIN, X. (1999). Modelling Labour Force Careers for the LifePaths Simulation Model, *Proceedings: Symposium 99 Combining Data from Different Sources*, Statistics Canada, 57-64.
- WOLFSON, M.C. (1997). Sketching LifePaths: A New Framework for Socio-Economic Statistics. *Simulating Social Phenomena*, (Eds. Conte, R. Gegselmann and P. Terna), Lecture Notes in Economics and Mathematical Systems, 456, Springer.
- WOLFSON, M.C., and ROWE, G. (1996). Perspectives on Working Time Over the Life Cycle, Canadian Employment Research Forum Conference on Changes to Working Time, Ottawa.
- WOLFSON, M.C., and ROWE, G. (1998a). LifePaths - Toward an Integrated Microanalytic Framework for Socio-Economic Statistics. 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.
- WOLFSON, M.C., and ROWE, G. (1998b). Public Pension Reforms - Analyses Based on the LifePaths Generational Accounting Framework, 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.
- WOLFSON, M.C., ROWE, G., GRIBBLE, S. and LIN, X. (1998). Historical Generational Accounting with Heterogeneous Populations. *Government Finances and Generational Equity* (Ed. M. Corak), Statistics Canada, 107-127.

Improved Ratio Estimation in Telephone Surveys Adjusting for Noncoverage

STEVEN T. GARREN and TED C. CHANG¹

ABSTRACT

Since some individuals in a population may lack phones, telephone surveys using random digit dialing within strata may result in asymptotically biased estimators of ratios. The impact from not being able to sample the nonphone population is examined. We take into account the propensity that a household owns a phone, when proposing a post-stratified phone-weighted estimator, which seems to perform better than the typical post-stratified estimator in terms of mean squared error. Such coverage propensities are estimated using the Public Use Microdata Samples, as provided by the United States Census. Non-post-stratified estimators are considered when sample sizes are small. The asymptotic mean squared error, along with its estimate based on a sample, of each of the estimators is derived. Real examples are analyzed using the Public Use Microdata Samples. Other forms of nonresponse are not examined herein.

KEY WORDS: Asymptotics; Census Public Use Microdata Samples; Post-stratification; Telephone survey.

1. INTRODUCTION

Consider surveys where the telephone population is sampled. Major problems in telephone surveys include nonresponse (*i.e.*, refusal to participate in the survey) and noncoverage (*i.e.*, lacking telephone service). Nonresponse may cause larger bias than noncoverage, since nonresponse propensities are usually much higher than noncoverage propensities. However, nonresponse is reviewed rather briefly, because the focus of this article is noncoverage.

1.1 Literature Review

Khurshid and Sahai (1995) provided an extensive bibliography of papers on telephone surveys. Examples of nonresponse rates may be found in Steeh, Groves, Comment and Hansmire (1983, pages 189-197). Corrections for nonresponse, using weights and imputation, were discussed by Little (1986) and Rubin (1987). Rao (1997) provided an overview of sample surveys, including discussions on resampling methods, especially the jackknife, for variance estimation. His discussion includes techniques to estimate the variance in the presence of imputation.

Regarding noncoverage, Brick, Waksberg and Keeter (1994) found the 94% of the households in the United States have phones at any given time. They also found that the households with interrupted telephone service usually are indigent. Keeter (1995) discussed that in a survey conducted from 1992 to 1993 more than half of all households without continuous telephone service during that year were *transient*, *i.e.*, these transient households were both with and without telephone service at different times during that year. He also found that most socioeconomic factors

(excluding home ownership) for transient telephone households are similar to those factors for households which are continuously without phones. These similarities between the transient and the nonphone populations suggest that valid inferences may be made on the entire (phone, nonphone, and transient) population, based on telephone surveys. Thornberry and Massey (1988) examined noncoverage for various socio-demographic groups from 1963 to 1986, and found income to be the most important factor in determining the likelihood that a household has a phone.

1.2 Our Approach

Given several various characteristics, such as home ownership and household language, the propensity of a household to have phone service is estimated in this article using the Virginia portion of the 1990 Census Public Use Microdata Samples (PUMS), which represent 5% of the population. Whether or not households have phones is included in the PUMS. The estimation of these propensities, or probabilities of phone service, is based on generalized linear regression with a log – log link, since the logit link provides a poor fit. We advocate using our fitted regression model, with the estimated parameters, for estimating these likelihoods in general whenever a random sample is taken from the Virginian phone population.

Because it is such a huge data set, the PUMS have another useful purpose in this article. The PUMS are used to compare and contrast estimators in terms of bias and variance, by examining the entire phone population and by taking repeated samples of the phone population. Categorical data consisting of 75 household and 75 personal variables are listed for all individuals in the households selected to be in the PUMS.

¹ Steven T. Garren, Department of Mathematics and Statistics, MSC 7803, James Madison University, Harrisonburg, Virginia, 22807, U.S.A. Research partially supported by NIMH grant MH53259-01A2; Ted C. Chang, Division of Statistics, 108 Halsey Hall, University of Virginia, Charlottesville, Virginia, 22903, U.S.A. Research partially supported by ONR grant N-00014-92-J-1009.

In the examples in section 6 high school graduation rate, mean number of cars per household, and mean household income are estimated using both post-stratified and non-post-stratified estimators for samples of size 500 from the PUMS. The post-stratification variables for high school graduation rate are gender, age, and race of the head of household. The post-stratification variable for mean number of cars per household is household income only. Estimators of the mean household income are analysed twice. For one analysis, post-stratification is on only the race of the head of household. For the other analysis, post-stratification is on gender, age, and race of the head of household. Each of these post-stratification variables is divided into two categories, except income, which is divided among three categories.

A serious drawback to estimators not taking into account the propensities of phone service is that these estimators are not asymptotically unbiased as the sample size gets large. A major focus of this article is to show that bias is reduced substantially when the estimators take into account the propensities of phone service, as estimated by the PUMS. Since both *post-stratified* and *non-post-stratified* estimators as well as both *using* and *not using* the propensities of phone service are considered, then four estimators are examined herein. In particular, these four estimators of a population mean are the sample mean, the usual post-stratified estimator, a phone-weighted estimator, and a proposed post-stratified phone-weighted estimator. The mean squared errors (MSE) of the phone-weighted estimator and the post-stratified phone-weighted estimator go to zero as the sample size gets large, unlike the other two estimators.

We adopt a two-phase model for our four estimators. The first phase involves selection from the entire population into the phone population. We treat the propensity of a household to have phone service as the probability that the household will be selected into the phone population, and we assume that this probability is positive (although possibly small) for each household. The second phase is a stratified (perhaps geographically stratified) simple random sample from the phone population. In the examples in section 6, we consider post-stratification by characteristics such as race and age of the head of household. Since our sample sizes are small, we do not geographically stratify the population of Virginia, although our formulas allow for both stratification and post-stratification.

Ideally, one would post-stratify using the same covariates used for estimating the propensities of phone service in the first phase of our model. In this case, the three estimators which use the propensities of phone service and/or post-stratification will be almost identical. However, the sample size for each post-stratified category should not be too small, so practical limitations restrict the number of categories which should be used for post-stratification. Nevertheless, many categories may be used for constructing the propensities of phone service from the PUMS, because the entire population is used.

Even if post-stratification by many covariates is feasible, the usual variance formulas for post-stratification require that a stratified random sample be taken from the entire population. In our situation, however, a stratified random sample is taken from the phone population, so the usual variance formulas are not applicable to our situation. The techniques by Politz and Simmons (*cf.* Cochran 1977, pages 374-377) require the sampling frame to be the entire population, not just the phone population, and hence are not applicable to our scenario, which allows noncoverage.

We derive the asymptotic variances of the four estimators of a population ratio, and we determine reasonable estimates of these variances. Since a population mean is a special case of a ratio, and a population total is a multiple of a ratio, then the results regarding estimators of means or totals follow from the results regarding estimators of ratios.

2. NOTATION

Consider N households in a population, U . For each household in U , let two variables of interest be denoted by y_{1k} and y_{2k} , for $k \in U$. At any given time, the event that the k th household does or does not have a phone is treated as random, while y_{1k} is treated as fixed.

Letting

$$\alpha_i = N^{-1} \sum_{k \in U} y_{ik},$$

for $i = 1, 2$, the goal is to estimate α_1 , α_2 , and the ratio

$$\mu = \alpha_1 / \alpha_2.$$

Without loss of generality we concentrate on estimating α_1 and μ .

An important special case of estimating a ratio μ arises when one desires to estimate the mean of a variable z_k for $k \in D$ for some subpopulation $D \subset U$ but one cannot sample directly from D . Examples include subpopulations defined by race. Let x_k be 1 if $k \in D$ and 0 otherwise. Let $y_{1k} = z_k x_k$ and $y_{2k} = x_k$. Then μ is the population mean of z_k over the subpopulation D .

Assume there are H strata, and h is used to index the strata. Assume there are G groups, and g is used to index the groups, which are used to construct the post-strata. The strata are known prior to sampling, but the groups are not observed until after the final sample is taken. Therefore, U_{gh} denotes all households in group g and stratum h ; N_{gh} denotes the size of U_{gh} , and N_h denotes the size of U_h . Other variables are defined similarly in terms of g and h .

Let U_T denote the population of households in U which currently have telephones, and let N_T denote the size of U_T . The probability, or propensity, that the k th household in U is also in U_T is denoted by p_k , and we assume that $p_k > 0$ for all k . A simple random sample of size n_h is taken from U_{Th} for $h = 1, \dots, H$. Let s_h denote this final sample in stratum h . The size of the final sample, s , is denoted by

n . For asymptotics herein, we assume that $n/N \rightarrow 0$ as $n \rightarrow \infty$ in the same spirit as Särndal, Swensson and Wretman (1992, pages 166-169).

3. THE ESTIMATORS

The sampling design is treated as a two-phase design with Poisson sampling at the first phase and stratified simple random sampling at the second phase. Each individual enters the telephone population with probability p_k , for $k \in U$, and then enters the final sample according to a simple random sample of size n_h , $h = 1, \dots, H$. The p_k are assumed known or can be estimated accurately, as shown in section 5. The estimators of μ discussed in this section will be validated in the appendix.

3.1 The Post-stratified and Ratio Estimators

Post-stratified estimates of α_1 and α_2 are

$$\hat{\alpha}_{ps(i)} = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} n_{gh}^{-1} \sum_{k \in s_{gh}} y_{ik},$$

for $i = 1, 2$, and the post-stratified estimate of μ is $\hat{\mu}_{ps} = \hat{\alpha}_{ps(1)} / \hat{\alpha}_{ps(2)}$. A valid estimate of the variance, conditional on U_T , is known to be (cf. Särndal *et al.* 1992, pages 270-271)

$$\begin{aligned} \widehat{\text{var}}(\hat{\mu}_{ps} | U_T) \\ = (N \hat{\alpha}_{ps(2)})^{-2} \sum_{h=1}^H \sum_{g=1}^G N_{gh}^2 \left[\frac{1 - (n_h / N_{Th})}{n_{gh}(n_{gh} - 1)} \right] \\ \sum_{k \in s_{gh}} \left[y_{1k} - \hat{\mu}_{ps} y_{2k} - n_{gh}^{-1} \sum_{j \in s_{gh}} (y_{1j} - \hat{\mu}_{ps} y_{2j}) \right]^2. \end{aligned} \quad (3.1)$$

Although the bias cannot be estimated from the final sample, the theoretical bias of $\hat{\mu}_{ps}$ is well-known to be

$$\begin{aligned} \text{bias } \hat{\mu}_{ps} = \frac{\sum_{h=1}^H \sum_{g=1}^G N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{1k}}{\sum_{h=1}^H \sum_{g=1}^G N_{gh} \left[\sum_{j \in U_{gh}} p_j \right]^{-1} \sum_{k \in U_{gh}} p_k y_{2k}} \\ - \frac{\sum_{k \in U} y_{1k}}{\sum_{k \in U} y_{2k}} + O(n^{-1}) \end{aligned} \quad (3.2)$$

as $n \rightarrow \infty$. Noting (3.2), the MSE of $\hat{\mu}_{ps}$ does not go to zero in general as the sample size n gets large.

To determine the variance and bias of $\hat{\alpha}_{ps(1)}$, set $y_{2k} = 1$ for all k , so that $\hat{\mu}_{ps}$ and μ become $\hat{\alpha}_{ps(1)}$ and α_1 , respectively. One may then apply (3.1) and (3.2) so that

$$\begin{aligned} \widehat{\text{var}} \hat{\alpha}_{ps(1)} = N^{-2} \sum_{h=1}^H \sum_{g=1}^G N_{gh}^2 \left[\frac{1 - (n_h / N_{Th})}{n_{gh}(n_{gh} - 1)} \right] \\ \sum_{k \in s_{gh}} \left[y_{1k} - n_{gh}^{-1} \sum_{j \in s_{gh}} y_{1j} \right]^2, \end{aligned} \quad (3.3)$$

and

$$\begin{aligned} \text{bias } \hat{\alpha}_{ps(1)} \\ = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} \left(\sum_{j \in U_{gh}} p_j \right)^{-1} \sum_{k \in U_{gh}} p_k y_{1k} - \alpha_1 + O(n^{-1}) \\ = O(1) \end{aligned}$$

as $n \rightarrow \infty$. Cochran (1977, pages 134-135) provided a correction factor, which is of order n^{-2} , to (3.3). This correction factor, however, is irrelevant to (3.1), since the error term due to estimation from the ratio is $O(n^{-2})$.

As usual, the *ratio estimator*, denoted by \bar{y}_1 / \bar{y}_2 , is defined to be the ratio of the sample mean of y_1 to the sample mean of y_2 . That is,

$$\bar{y}_1 / \bar{y}_2 = \sum_{k \in s} y_{1k} / \sum_{j \in s} y_{2j}.$$

The post-stratified and ratio estimators are identical when $G = H = 1$. Since we will be using only one stratum (i.e., $H = 1$) in section 6, we need not reference separate theory for the ratio estimator.

3.2 The Phone-weighted Estimator

Since the post-stratified estimator, $\hat{\mu}_{ps}$, is biased, two alternative estimators are suggested. One is the phone-weighted estimator, which takes into account the probability that an individual has a phone. In this section we assume that the p_k are known for all $k \in s$ or can be estimated accurately. Estimation of p_k using the PUMS is discussed in section 5.

For a crude estimate of α_i for $i = 1, 2$, use

$$\tilde{\alpha}_{w(i)} = N^{-1} \sum_{h=1}^H N_{Th} n_h^{-1} \sum_{k \in s_h} p_k^{-1} y_{ik}. \quad (3.4)$$

Then, estimate μ by

$$\hat{\mu}_w = \tilde{\alpha}_{w(1)} / \tilde{\alpha}_{w(2)}, \quad (3.5)$$

which is asymptotically unbiased for μ , since $\tilde{\alpha}_{w(i)}$ is unbiased for $\alpha_{w(i)}$, for $i = 1, 2$. A valid estimate of the variance of $\hat{\mu}_w$ is shown to be

$$\widehat{\text{var}} \hat{\mu}_w = [N \tilde{\alpha}_{w(2)}]^{-2} \sum_{h=1}^H \frac{N_{Th}^2 [1 - (n_h/N_{Th})]}{n_h(n_h - 1)} \sum_{k \in s_h} \left[\frac{y_{1k} - \hat{\mu}_w y_{2k}}{p_k} - n_h^{-1} \sum_{j \in s_h} \frac{y_{1j} - \hat{\mu}_w y_{2j}}{p_j} \right]^2. \quad (3.6)$$

Since the estimator, $\hat{\mu}_w$, is asymptotically unbiased, then a valid estimate of the MSE of $\hat{\mu}_w$ is identical to the estimate of the variance.

Setting $y_{2j} = 1$ in (3.4) and (3.5) allows a valid estimate of $\alpha_{w(1)}$ to be

$$\tilde{\alpha}_{w(1)} = \left[\sum_{h=1}^H N_{Th} n_h^{-1} \sum_{k \in s_h} p_k^{-1} \right]^{-1} \sum_{h=1}^H N_{Th} n_h^{-1} \sum_{k \in s_h} p_k^{-1} y_{1k}.$$

The variance of $\hat{\alpha}_{w(1)}$ may be estimated by setting $y_{2j} = 1$ in (3.6).

3.3 The Post-stratified Phone-weighted Estimator

Another proposed estimator combines post-stratification with the phone-weighted estimator, and is perhaps the best among the four, when sample sizes are large enough to justify post-stratification. This new estimator requires, however, that all N_{gh} be large enough so that with high probability the n_{gh} are not too small. To estimate α_i we use

$$\tilde{\alpha}_{\text{psw}(i)} = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} \left[\sum_{j \in s_{gh}} p_j^{-1} \right]^{-1} \sum_{k \in s_{gh}} p_k^{-1} y_{ik},$$

for $i = 1, 2$. We then estimate μ by $\hat{\mu}_{\text{psw}} = \hat{\alpha}_{\text{psw}(1)} / \hat{\alpha}_{\text{psw}(2)}$. The estimate of the variance of $\hat{\mu}_{\text{psw}}$ is

$$\widehat{\text{var}} \hat{\mu}_{\text{psw}} = [N \tilde{\alpha}_{\text{psw}(2)}]^{-2} \sum_{h=1}^H \sum_{g=1}^G N_{gh}^2 \left(\sum_{j \in s_{gh}} p_j^{-1} \right)^{-2} \left(\frac{n_{gh}}{n_{gh} - 1} \right) \left(1 - \frac{n_h}{N_{Th}} \right) \sum_{k \in s_{gh}} p_k^{-2} \left[y_{1k} - \hat{\mu}_{\text{psw}} y_{2k} - \left(\sum_{j \in s_{gh}} p_j^{-1} \right)^{-1} \sum_{m \in s_{gh}} p_m^{-1} (y_{1m} - \hat{\mu}_{\text{psw}} y_{2m}) \right]^2. \quad (3.7)$$

If any of the n_{gh} terms are small, then one might instead prefer the estimator

$$\widehat{\text{var}} \hat{\mu}_{\text{psw}} = [N \tilde{\alpha}_{\text{psw}(2)}]^{-2} \sum_{h=1}^H N_h^2 \left(\sum_{j \in s_h} p_j^{-1} \right)^{-2} \left(\frac{n_h}{n_h - 1} \right) \left(1 - \frac{n_h}{N_{Th}} \right) \sum_{g=1}^G \sum_{k \in s_{gh}} p_k^{-2} \left[y_{1k} - \hat{\mu}_{\text{psw}} y_{2k} - \left(\sum_{j \in s_{gh}} p_j^{-1} \right)^{-1} \sum_{m \in s_{gh}} p_m^{-1} (y_{1m} - \hat{\mu}_{\text{psw}} y_{2m}) \right]^2. \quad (3.8)$$

Notice that if N_{Tgh} were known, which is however unlikely, then a more familiar and intuitive estimator of $\text{var} \hat{\mu}_{\text{psw}}$ would be

$$\widehat{\text{var}} \hat{\mu}_{\text{psw}} = [N \tilde{\alpha}_{\text{psw}(2)}]^{-2} \sum_{h=1}^H \sum_{g=1}^G \frac{N_{Tgh}^2}{n_{gh}(n_{gh} - 1)} \left(1 - \frac{n_{gh}}{N_{Tgh}} \right) \sum_{k \in s_{gh}} p_k^{-2} \left[y_{1k} - \hat{\mu}_{\text{psw}} y_{2k} - \left(\sum_{j \in s_{gh}} p_j^{-1} \right)^{-1} \sum_{m \in s_{gh}} p_m^{-1} (y_{1m} - \hat{\mu}_{\text{psw}} y_{2m}) \right]^2. \quad (3.9)$$

Since N_{Tgh} typically is unknown, then (3.9) usually is not a practical estimator. However, (3.9) helps motivate (3.7) and (3.8), which are quite practical.

Since the estimator, $\hat{\mu}_{\text{psw}}$, is asymptotically unbiased, then a valid estimate of the MSE is identical to the estimate of the variance. Further, setting $y_{2j} = 1$ in (3.7) and (3.8) allows one to estimate the variance of $\hat{\alpha}_{\text{psw}(1)}$.

When $G = 1$, the estimator $\hat{\mu}_{\text{psw}}$ does not reduce to $\hat{\mu}_w$, as one might naively anticipate. The preferred estimator when $G = 1$ is $\hat{\mu}_w$, since $\hat{\mu}_w$ is based on only one ratio, whereas $\hat{\mu}_{\text{psw}}$ is based on a ratio of ratios. The estimator $\hat{\mu}_{\text{psw}}$ requires large sample sizes in each stratum-group category, but $\hat{\mu}_w$ requires only a large overall sample size. When $H = G = 1$, however, the estimators $\hat{\mu}_w$ and $\hat{\mu}_{\text{psw}}$ are identical; the variance estimators based on $\hat{\mu}_w$ are preferable to those based on $\hat{\mu}_{\text{psw}}$ because the estimates of the variance of $\hat{\mu}_{\text{psw}}$ are based on a ratio of ratios.

4. ASYMPTOTIC MEAN SQUARED ERRORS

The asymptotic mean squared errors of the estimators defined in section 3 now are stated. The proofs follow from Taylor linearization and are given in the appendix, along with the minor regularity conditions needed.

4.1 The Post-stratified Estimator

To find the asymptotic theoretical variance of the post-stratified estimator of μ , we first define

$$\alpha_i^* = \lim_{n \rightarrow \infty} \hat{\alpha}_{ps(i)} = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} \left(\sum_{j \in U_{gh}} p_j \right)^{-1} \sum_{k \in U_{gh}} p_k y_{ik}, \quad (4.1)$$

for $i = 1, 2$, and also define

$$\mu^* = \alpha_1^* / \alpha_2^*. \quad (4.2)$$

Note that $\alpha_i^* \neq \alpha_i$ and $\mu^* \neq \mu$ in general. The asymptotic theoretical variance of $\hat{\mu}_{ps}$ is

$$\begin{aligned} \text{var } \hat{\mu}_{ps} &= (N\alpha_2^*)^{-2} \sum_{h=1}^H \sum_{g=1}^G \\ &\frac{N_{gh}^2 \left[\left(\sum_{j \in U_{gh}} p_j \right) - n_h \right]}{n_h \left(\sum_{j \in U_{gh}} p_j \right) \left[\left(\sum_{j \in U_{gh}} p_j \right) - 1 \right]} \sum_{k \in U_{gh}} p_k \\ &\left[y_{1k} - \mu^* y_{2k} - \frac{\sum_{j \in U_{gh}} p_j (y_{1j} - \mu^* y_{2j})}{\sum_{j \in U_{gh}} p_j} \right]^2 + O(n^{-2} + N^{-1}) \end{aligned} \quad (4.3)$$

as $n \rightarrow \infty$. The asymptotic bias of $\hat{\mu}_{ps}$ was shown in (3.2) to be $O(1)$ as $n \rightarrow \infty$. Therefore, the asymptotic MSE of $\hat{\mu}_{ps}$ is also $O(1)$ as $n \rightarrow \infty$.

4.2 The Phone-weighted Estimator

The asymptotic theoretical variance of the phone-weighted estimator of μ is

$$\begin{aligned} \text{var } \hat{\mu}_w &= (N\alpha_2)^{-2} \sum_{h=1}^H \frac{\left[\left(\sum_{j \in U_h} p_j \right) - n_h \right] \left(\sum_{j \in U_h} p_j \right)}{n_h \left[\left(\sum_{j \in U_h} p_j \right) - 1 \right]} \\ &\sum_{k \in U_h} p_k \left[\frac{y_{1k} - \mu y_{2k}}{p_k} - \frac{\sum_{j \in U_h} (y_{1j} - \mu y_{2j})}{\sum_{j \in U_h} p_j} \right]^2 + O(n^{-2} + N^{-1}). \end{aligned} \quad (4.4)$$

Since $\hat{\mu}_w$ is asymptotically unbiased, then its MSE is the same as the right hand side of (4.4).

4.3 The Post-stratified Phone-weighted Estimator

The asymptotic theoretical variance of the post-stratified phone-weighted estimator of μ is

$$\begin{aligned} \text{var } \hat{\mu}_{psw} &= (N\alpha_2)^{-2} \sum_{h=1}^H \sum_{g=1}^G \\ &\frac{\left(\sum_{j \in U_{gh}} p_j \right) \left[\left(\sum_{j \in U_{gh}} p_j \right) - n_h \right]}{n_h \left[\left(\sum_{j \in U_{gh}} p_j \right) - 1 \right]} \sum_{k \in U_{gh}} p_k^{-1} \\ &\left[y_{1k} - \mu y_{2k} - N_{gh}^{-1} \sum_{j \in U_{gh}} (y_{1j} - \mu y_{2j}) \right]^2 \\ &+ O(n^{-2} + N^{-1}). \end{aligned} \quad (4.5)$$

Since $\hat{\mu}_{psw}$ is asymptotically unbiased, then its MSE is the same as the right hand side of (4.5).

5. ESTIMATING THE p_k USING PUBLIC USE MICRODATA SAMPLES

The United States Bureau of the Census produced the Public Use Microdata Samples (PUMS), which include 1% and 5% samples of the population in each of the 50 states and Washington, D.C., for year 1990. For each person selected in the sample, 75 household variables and 75 personal variables are listed, where each household has a clearly defined head of household. We utilize the PUMS for two reasons. We estimate the p_k using the PUMS in this section, whereas in section 6 we run simulations on the PUMS to construct examples for comparing and contrasting the estimators.

In this article, we use the 5% sample from Virginia. Since 5% represents a huge number of households, we treat this sample as if it were the entire population of Virginia. Since we are interested in telephone surveys, then from this 5% sample we will sample households. Inferences may be made on personal variables, such as high school graduation rate, and household variables, such as the number of cars in a household or household income. Information pertaining to whether or not each household has a phone is included in the PUMS. We removed from our study all households whose telephone status is listed as "not applicable." Such households were either vacant or were group quarters (institutions and non-institutions). The number of households remaining in 1990 is 110,744, of which 104,606 have phones; hence, the proportion of these households which have phones is 94.5%.

Using generalized linear regression with a log-log link on the 5% sample from Virginia along with the household weights assigned in the PUMS, we estimate p_k , which is the probability, or propensity, that the k th household has a phone. McCullagh and Nelder (1991, pages 107-110)

recommended the use of a log – log link when the probabilities are close to one, and we found that this link provided a good fit. We also found that the logit link function provided a poor fit.

The PUMS household weights are used when estimating the p_k but are not used elsewhere herein. In particular, in section 6 when constructing Monte Carlo samples of the PUMS population, the samples are simple random samples from the telephone population.

Examples of estimating the p_k

Six covariates, the number of persons in the household, tenure (home owner or renter), the date the head of household moved into the dwelling, household income, household language, and race of the head of household, are used to estimate the p_k . These six covariates were chosen, along with the categories for each covariate, based on a thorough analysis of the 1990 PUMS using generalized linear regression techniques in SAS. All of these covariates were found to be highly statistically significant. Estimates of the p_k are made by summing the appropriate estimates of the covariates in Table 1. The covariate for the number of persons should be multiplied by the number of persons in the household; however, if the number of persons exceed

five, then, for computations, convert this number of persons to five. For example, if the household consists of three English-speaking Asian Americans with two cars in a house purchased in 1987, where the household income is \$75,000, then Table 1 indicates that the estimate of p_k is the solution to

$$\log(-\log p_k) = 3 \times 0.2747 - 0.5552 + 0.5920 \\ + 0.1896 + 1.0004 + 0.6156 + 0.0000.$$

Notice that in Table 1 within each of the covariates *date moved in*, *number of cars*, and *income*, the values corresponding to the categories are monotonically decreasing, as anticipated, except when income is negative.

An adjustment which should be made when using random digit dialing is to ask each respondent the number of phone lines in the household, and multiply that number by the estimate of p_k from Table 1 to obtain a new estimate of p_k . Consequently, p_k now is a weight, rather than a probability. For the simulations discussed in section 6 this adjustment is not necessary, since households are equally likely to be selected using simple random sampling from the PUMS, regardless of the number of phone lines.

Table 1

Values of covariates for estimating p_k using the Virginia 5% PUMS. Standard errors are in parentheses. If the number of persons exceeds five, then convert this number to five. The covariate "tenure" did not appear in the 1980 PUMS. The 1980 category "\$40,000 to \$49,999" actually includes "\$40,000 or greater". The "other" category for the 1980 covariate "language" includes Spanish.

Covariate	Category	1990 Value		1980 Value	
Number of persons		0.2747	(0.0022)	0.1929	(0.0020)
tenure	home owner	-0.5552	(0.0079)	-0.7845	(0.0057)
	renter	0.0000	(0.0000)	0.0000	(0.0000)
date moved in	1989 or 1990	0.9742	(0.0121)	NA	
	1985 to 1988	0.5920	(0.0119)	NA	
	1980 to 1984	0.3489	(0.0138)	NA	
	1970 to 1979	0.2185	(0.0136)	NA	
	1969 or earlier	0.0000	(0.0000)	NA	
number of cars	0	1.2927	(0.0152)	0.8633	(0.0118)
	1	0.6842	(0.0143)	0.3981	(0.0109)
	2	0.1896	(0.0145)	0.0399	(0.0112)
	3 or more	0.0000	(0.0000)	0.0000	(0.0000)
income	less than \$0	3.5325	(0.1294)	2.3639	(0.0830)
	\$0 to \$9,999	3.7929	(0.0539)	2.5238	(0.0260)
	\$10,000 to \$19,999	3.4878	(0.0538)	1.9763	(0.0258)
	\$20,000 to \$29,999	3.0299	(0.0539)	1.0220	(0.0269)
	\$30,000 to \$39,999	2.4297	(0.0543)	0.3889	(0.0317)
	\$40,000 to \$49,999	1.8899	(0.0556)	0.0000	(0.0000)
	\$50,000 to \$59,999	1.5992	(0.0578)	NA	
	\$60,000 to \$69,999	1.2144	(0.0631)	NA	
	\$70,000 to \$79,999	1.0004	(0.0704)	NA	
	\$80,000 or greater	0.0000	(0.0000)	NA	
language	English	0.6156	(0.0164)	0.4232	(0.0153)
	Spanish	0.4889	(0.0216)	NA	
	other	0.0000	(0.0000)	0.000	(0.0000)
race	black	-0.4233	(0.0064)	-0.3837	(0.0058)
	other	0.0000	(0.0000)	0.0000	(0.0000)
intercept		-7.6707	(0.0588)	-4.9024	(0.0322)

Table 1 thus can be used for estimating p_k when conducting telephone surveys. When a generalized linear regression model calculated from a PUMS of an earlier date is used to analyse a later survey, rescaling should be performed to take into account changes in the distribution of household income across time. Table 1 also gives the coefficients of a model calculated from the 1980 PUMS. We discuss in section 6 an example when the 1980 PUMS model is used to calculate p_k for a sample from the 1990 PUMS population. We note that the 1980 PUMS did not include "date moved in" and that a better fitting model arose when the language categories "Spanish" and "other" were combined. In addition, median household income almost doubled between 1980 and 1990, so fewer income categories were used in 1980.

Although Table 1 is convenient when sampling from the PUMS and performing simulations, the covariates listed in Table 1 might not be available in actual surveys involving random digit dialing. One may reproduce Table 1 using different covariates, or one may estimate the p_k according to the following alternative method.

An alternative method for estimating the p_k

The participants in a telephone survey based on random digit dialing may be asked the following two questions: "(1) How many telephone lines have been in your household during the past twelve months? (2) During the past twelve months, how many months was each telephone line in service?" Now, let p_k be the sum of the answers to question (2). For example, in a household with two phone lines, where one of the lines was in service all twelve months and the other was in service only five months, the estimate of p_k would be $12+5=17$. Again, p_k represents a weight rather than a probability here. Asking the respondent this second question is similar to an approach advocated by Brick *et al.* (1994), who also suggested weighting the data to take into account the probability that a household has phone service.

6. INFERENCES ON HOUSEHOLD AND PERSONAL VARIABLES

We will compare the four proposed estimators of μ as we make inferences on the high school graduation rate among people at least 21-years-old, the mean number of cars per household, and the mean household income, in the state of Virginia. We performed 100,000 simulations of simple random samples of 500 households with telephones from the 1990 Virginia 5% PUMS using one stratum (*i.e.*, $H=1$).

In section 6.1, two sets of p_k are used. One is based upon a GLIM regression fit to the 1990 PUMS, and the other is based upon a GLIM fit to the 1980 PUMS with the income categories inflated by the ratio of the 1990 median household income (\$32,800) to the 1980 median household income (\$17,510). Using the 1980 p_k to estimate a 1990 parameter demonstrates how well our method works when

GLIM coefficients are used for future data sets, provided than an adjustment for inflation is made. Only the 1990 p_k are used in section 6.2 and section 6.3.

Post-stratification should be used when the sample sizes are sufficiently large. Non-post-stratified estimators may be compared to each other, and post-stratified estimators may be compared to each other. Comparing \bar{y}_1/\bar{y}_2 to $\hat{\mu}_w$ is appropriate, and comparing $\hat{\mu}_{ps}$ to $\hat{\mu}_{psw}$ is appropriate. These comparisons show the improvements when using the p_k in the estimators.

6.1 Estimating the High School Graduation Rate

Using the entire 1990 Virginia 5% PUMS, the mean high school graduation rate among all Virginians at least 21-years-old is $\mu=0.75118$. When estimating the graduation rate using a simple random sample and $\hat{\mu}_{ps}$ or $\hat{\mu}_{psw}$, we post-stratify by gender (male, female), age (less than 45 years old, at least 45 years old), and race (black, other) of the head of household. The p_k are estimated using Table 1. The values of the biases and standard deviations discussed below are shown in Table 2, when 1990 p_k are used.

Table 2
Biases and Standard Deviations of Estimates of High School Graduation Rate

	Estimator			
	not post-stratified	post-stratified		
	\bar{y}_1/\bar{y}_2	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	0.01471	0.00722	0.01461	0.00874
telephone bias	0.01472	0.00720	0.01463	0.00850
second phase bias	0.00000	0.00002	-0.00002	0.00024
theoretical bias	0.00777	0	0.00663	0
simulated standard deviation	0.01683	0.01737	0.01605	0.01643
estimated standard deviation	0.01680	0.01734	0.01601	0.01635*
theoretical standard deviation	0.01700	0.01752	0.01617	0.01658
root mean squared error	0.02236	0.01881	0.02171	0.01861

The true high school graduation rate is 0.75118. Post-stratification is based on gender, age, and race. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is 0.01610.

The *aggregate biases* of the four estimators of μ are estimated by the average over 100,000 simulations of the difference between the estimate from a sample of size 500 and μ . These *aggregate biases* produced by \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are estimated to be 0.01471, 0.00722, 0.01461, and 0.00874, respectively, when 1990 p_k are used. Hence using the p_k reduces the bias of the non-post-stratified estimator by 51%, and reduces the bias of the post-stratified estimator by 40%.

When the 1980 p_k are used, similar results arise. These *aggregate biases* produced by $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are estimated to be 0.00578, and 0.00856, respectively, when the 1980 p_k are used. These results, however, are not summarized in the tables.

The *telephone bias*, listed in Table 2, is the bias obtained when the entire telephone population, U_T , is sampled when calculating \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$. This bias is caused by the fact that U_T is sampled rather than U . Throughout this example, we use the convention of listing the estimates based on the 1980 p_k in parentheses, when these estimates differ from those based on the 1990 p_k . The *telephone biases* are 0.01472, 0.00720 (0.00577), 0.01463, and 0.00850 (0.00838), and are relatively close to the *aggregate biases*.

The *second phase bias* is the difference between the *aggregate bias* and the *telephone bias*, and is caused by the fact that the estimator approximates a ratio. This *second phase bias*, modulus rounding error, for \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are estimated to be 0.00000, 0.00002 (0.00001), -0.00002, and 0.00024 (0.00018), respectively. Hence, the *second phase bias* is trivial compared to the *telephone bias* for this example.

The *theoretical biases*, based on (3.2), of \bar{y}_1/\bar{y}_2 , and $\hat{\mu}_{ps}$ are 0.00777 (0.00905) and 0.00663 (0.00678), respectively. These biases differ from the *aggregate biases*, since (3.2) is based on all possible phone populations, whereas the *aggregate biases* are conditional on the one realization of the phone population. The theoretical bias is based upon the model that each household has a phone with probability p_k and hence is dependent upon the model used to fit p_k . Since $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are asymptotically unbiased, then their *theoretical biases* are defined to be zero.

The *simulated standard deviations* of the 100,000 simulated estimates of μ for \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are 0.01683, 0.01737 (0.01734), 0.01605, and 0.01643 (0.01634). These four numbers are fairly close to the *estimated standard deviations*, which are the squareroot of the average estimated variance of the estimator of μ , based on (3.1), (3.6), and (3.7). Specifically, these *estimated standard deviations* are 0.01680, 0.01734 (0.01732), 0.01601, and 0.01635 (0.01628), respectively. The estimated alternative standard deviation, based on (3.8), of $\hat{\mu}_{psw}$ is 0.01610 (0.01606), which again is fairly close to the value 0.01635 (0.01628). The *theoretical standard deviations* are 0.01700 (0.01697), 0.01752 (0.01749), 0.01617 (0.01621), and 0.01658 (0.01653), based on the entire 1990 Virginia 5% PUMS and (4.3), (4.4), and (4.5). These *theoretical standard deviations* also are close to the other standard deviations calculated.

Using the p_k reduces the *aggregate bias* in the non-post-stratified estimator by 51% (61%), and in the post-stratified estimator by 40% (41%). The standard deviation, however, increases slightly. Using the *aggregate biases* and the *simulated standard deviations*, the root mean squared errors of the estimators \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$ are 0.02236

(0.02236) 0.01881 (0.01828), 0.02171 (0.02171), and 0.01861 (0.01844), respectively. Hence, using the p_k reduces the MSE in the non-post-stratified estimator by 29% (33%), and reduces the MSE in the post-stratified estimator by 27% (28%). Notice that there is little difference between \bar{y}_1/\bar{y}_2 and $\hat{\mu}_{ps}$ and between $\hat{\mu}_w$ and $\hat{\mu}_{psw}$, in terms of MSE. Therefore, post-stratification offers little improvement.

6.2 Estimating the Mean Number of Cars per Household

The mean number of cars per household is 1.80397, as determined by the entire 1990 Virginia 5% PUMS. Post-stratification was based upon household income, using categories {less than \$20,000, at least \$20,000 but less than \$35,000, and at least \$35,000}. The p_k are again estimated, but this time the covariate "numbers of cars" was excluded from the GLIM fit to the 1990 PUMS, since mean number of cars per household is what is being estimated.

As shown in Table 3, the estimates of the *aggregate biases* using 100,000 simulations of 500 simple random samples are 0.04872, 0.01629, 0.02226, and 0.01471, and the *telephone biases* are 0.04872, 0.01623, 0.02220, and 0.01458, for estimators \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$, respectively. Therefore, the *second phase biases* are rather small. Using the p_k reduces the bias from the non-post-stratified estimator by 67%, and reduces the bias from the post-stratified estimator by 34%. Perhaps the reason why this latter amount of bias that can be removed is smaller than the former is that income is a strong predictor of whether or not a household has a phone (cf. Groves 1989, pages 116-119; Thornberry and Massey 1988), and the post-stratification groups for determining $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$ are based on income.

Table 3
Biases and Standard Deviations of Estimates of Mean Number of Cars per Household

	Estimator			
	not post-stratified	post-stratified		
	\bar{y}_1/\bar{y}_2	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	0.04872	0.01629	0.02226	0.01471
telephone bias	0.04872	0.01623	0.02220	0.01458
second phase bias	0.00000	0.00006	0.00006	0.00013
theoretical bias	0.03388	0	0.00859	0
simulated standard deviation	0.04694	0.04764	0.04162	0.04172
estimated standard deviation	0.04682	0.04753	0.04148	0.04158*
theoretical standard deviation	0.04715	0.04791	0.04152	0.04161
root mean squared error	0.06765	0.05035	0.04720	0.04424

The true mean number of cars per household is 1.80397. Post-stratification is based on income. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is 0.04142.

The standard deviations of the simulations are 0.04694, 0.04764, 0.04162, and 0.04172, respectively. The root mean squared errors for the four estimators are approximately 0.06765, 0.05035, 0.04720, and 0.04424, respectively, so using the p_k reduces the MSE by 45% and 12% for non-post-stratification and post-stratification, respectively.

We also performed simulations, not summarized in the tables, where "number of cars" was retained for the GLIM fit to the 1990 PUMS. These *aggregate biases* for the estimators $\hat{\mu}_w$ and $\hat{\mu}_{psw}$ are 0.00116 and 0.00006, respectively, which are much smaller than 0.01629 and 0.01471, the respective *aggregate biases* when "number of cars" was removed from the GLIM fit. Furthermore, we feel that appropriate analysis requires removing the variable being studied (*i.e.*, number of cars) from the GLIM fit to the PUMS.

6.3 Estimating the Mean Household Income

The mean household income is \$40,187, as determined by the entire 1990 Virginia 5% PUMS. The p_k are again estimated, but this time the covariate "income" was excluded from the GLIM fit to the 1990 PUMS, since mean household income is what is being estimated.

In Table 4, when estimating household income using a simple random sample of size 500 and $\hat{\mu}_{ps}$ or $\hat{\mu}_{psw}$, we post-stratified only by the race (black, other) of the head of household. The estimates of the *aggregate biases* using 100,000 simulations are \$1,412, \$640, \$1,192, and \$633, and the *telephone biases* are \$1,414, \$640, \$1,193, and \$630, for estimators \bar{y}_1/\bar{y}_2 , $\hat{\mu}_w$, $\hat{\mu}_{ps}$, and $\hat{\mu}_{psw}$, respectively. Thus, the *second phase biases* are small relative to the *telephone biases*. Overall, using the p_k reduces the bias from the non-post-stratified estimator by 55%, and reduces the bias from the post-stratified estimator by 47%.

The standard deviations of the simulations are \$1,534, \$1,518, \$1,502, and \$1,488, respectively. Hence the root mean squared errors for the four estimators are approximately \$2,085, \$1,647, \$1,918, and \$1,617, respectively, so using the p_k reduces the MSE by 38% and 29% for non-post-stratification and post-stratification, respectively. The improvements from using post-stratification are more minor, according to the MSE criterion.

In Table 5, we again are estimating household income, but this time we post-stratify by gender (male, female), age (less than 45 years old, at least 45 years old), and race (black, other) of the head of household. Note that the non-post-stratified estimators are not affected by this new post-stratification. The estimates of the *aggregate biases* using 100,000 simulations are \$1,173 and \$757, and the *telephone biases* are \$1,177 and \$747 for the post-stratified estimators, $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. Again, the *second phase biases* are small relative to the *telephone biases*. Using the p_k reduces the bias from merely post-stratification by 35%.

The theoretical bias for the post-stratified estimator is \$463. The standard deviations of the simulations are \$1,445

and \$1,435, for estimators $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. The root mean squared errors are \$1,861 and \$1,622, for estimators $\hat{\mu}_{ps}$ and $\hat{\mu}_{psw}$, respectively. Hence, using the p_k reduces the MSE of the post-stratified estimator by 24%.

The MSE of $\hat{\mu}_{psw}$ is approximately the same in Table 4 and Table 5. However, the MSE of $\hat{\mu}_{ps}$ decreases somewhat from Table 4 to Table 5.

Table 4
Biases and Standard Deviations of Estimates of Household Income, Post-stratified by Race

	Estimator			
	not post-stratified	post-stratified		
	\bar{y}_1/\bar{y}_2	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	\$1,412	\$640	\$1,192	\$633
telephone bias	\$1,414	\$640	\$1,193	\$630
second phase bias	-\$2	\$0	-\$2	\$3
theoretical bias	\$789	\$0	\$586	\$0
simulated standard deviation	\$1,534	\$1,518	\$1,502	\$1,488
estimated standard deviation	\$1,537	\$1,521	\$1,506	\$1,491*
theoretical standard deviation	\$1,535	\$1,518	\$1,503	\$1,488
root mean squared error	\$2,085	\$1,647	\$1,918	\$1,617

The true mean household income is \$40,187. Note that \bar{y}_1/\bar{y}_2 and $\hat{\mu}_w$ are independent of post-stratification, so their results are identical to those in Table 5. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is \$1,490.

Table 5
Biases and standard deviations of estimates of household income, post-stratified by gender, age, and race

	Estimator			
	not post-stratified	post-stratified		
	\bar{y}_1/\bar{y}_2	$\hat{\mu}_w$	$\hat{\mu}_{ps}$	$\hat{\mu}_{psw}$
aggregate bias	\$1,412	\$640	\$1,173	\$757
telephone bias	\$1,414	\$640	\$1,177	\$747
second phase bias	-\$2	\$0	-\$4	\$10
theoretical bias	\$789	\$0	\$463	\$0
simulated standard deviation	\$1,534	\$1,518	\$1,445	\$1,435
estimated standard deviation	\$1,537	\$1,521	\$1,448	\$1,438*
theoretical standard deviation	\$1,535	\$1,518	\$1,440	\$1,430
root mean squared error	\$2,085	\$1,647	\$1,861	\$1,622

The true mean household income is \$40,187. Note that \bar{y}_1/\bar{y}_2 and $\hat{\mu}_w$ are independent of post-stratification, so their results are identical to those in Table 4. Samples of size 500 were taken and 100,000 simulations were performed.

* This value is based on (3.7), whereas the value based on (3.8) is \$1,421.

7. DISCUSSION

We have proposed here to use publicly available large data bases (e.g., the PUMS) to develop a model for the propensity p_k of a household to have a telephone. We have used, for Virginia in 1990, a GLIM model with a log – log link and predictor variables number of persons, tenure, date moved in, number of cars, household income, language, and race.

We have proposed to use the telephone weights p_k to reduce the bias of estimators due to noncoverage in telephone surveys. This bias can be expected to occur when the variable of interest is related to telephone ownership. The examples we have chosen are all variables of this type and hence the improvements using telephone weights are better than one would expect for variables with little relationship to telephone ownership.

The weights can be combined with post-stratification. We have found that the use of such telephone weights greatly reduces the bias of both non-post-stratified and post-stratified estimators.

Post-stratification requires a large enough sample size so that each post stratum has a negligible probability of being empty. Our experiments dealt with samples of size 500, and hence the number of post strata was relatively limited. Certainly, if one had a large enough sample so that one could post-stratify on the same predictor variables as used to develop the p_k , the use of telephone weights should offer negligible improvement over post-stratification. However, many nationwide telephone opinion polls use approximate sample sizes of 1,000, and we believe for these sample sizes, the use of telephone weights would offer a genuine improvement.

We have also reported results from using telephone weights developed from the 1980 PUMS on 1990 data, with categories related to household income adjusted for inflation. The results are comparable to those for telephone weights developed from the 1990 PUMS. Therefore, although PUMS data are produced only every ten years and might be as much as twelve years out of date, substantial reductions in the biases of telephone sampling can be made using propensity models derived from older PUMS data sets, provided that the categories are suitably adjusted for inflation.

Finally, the PUMS are divided by state and major metropolitan areas. This allows separate telephone-weighted models to be developed for major geographical units, and this would seem appropriate for large surveys.

ACKNOWLEDGEMENT

The authors are very thankful to an anonymous referee for many helpful suggestions.

APPENDIX: DERIVATIONS OF EQUATIONS

Before deriving the equations in section 3 and section 4, some regularity conditions must be assumed for sequences $\{\alpha_{i1}, \alpha_{i2}, \dots\}$, for $i = 1, 2$. Further, some lemmas must be proved. Then, the equations involving the estimators $\hat{\mu}_{ps}$, $\hat{\mu}_w$, and $\hat{\mu}_{psw}$ will be derived in the subsections below. Whenever the error variable ξ_k is introduced below, then $\xi_k = O_p(1)$ and $E(\xi_k)^2 = O(1)$ as $k \rightarrow \infty$. For simplicity (but slight abuse) of notation, the sequence $\{\xi_1, \xi_2, \dots\}$ will be allowed to be different across different equations.

Condition A: Each α_{ik} represents a sample mean of observations such that $E\alpha_{ik} - \alpha_i = O(k^{-1})$, $E|\alpha_{ik} - \alpha_i|^3 = O(k^{-3/2})$, and $\alpha_{ik} - \alpha_i = O_p(k^{-1/2})$ as $k \rightarrow \infty$ for $i = 1, 2$. Let $\mu_k = \alpha_{1k}/\alpha_{2k}$ for $k = 1, 2, \dots$

LEMMA A.1 Condition A implies that $E\mu_k - \mu = O(k^{-1})$ as $k \rightarrow \infty$.

PROOF: Define the function $f(\gamma_1, \gamma_2) = \gamma_1/\gamma_2$. By a Taylor series linear expansion,

$$\begin{aligned} \mu_k - \mu &= \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2 \\ &= (\alpha_{1k} - \alpha_1) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} + k^{-1} \xi_k \\ &= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu(\alpha_2)^{-2} + k^{-1} \xi_k. \end{aligned}$$

The result follows from Condition A.

Condition B: The sequence $\{\alpha_{i1}, \alpha_{i2}, \dots\}$ for $i = 1, 2$ satisfies

$$\begin{aligned} k^{1/2} \left[\begin{pmatrix} \alpha_{1k} \\ \alpha_{2k} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \right] &\xrightarrow{d} \\ N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) &+ \begin{pmatrix} k^{-1/2} \xi_{1k} \\ k^{-1/2} \xi_{2k} \end{pmatrix} \end{aligned}$$

for some constants σ_1^2 , σ_2^2 , and ρ .

LEMMA A.2 Under Conditions A and B,

$$\begin{aligned} \text{MSE } \mu_k &= (\alpha_2)^{-2} \text{var}(\alpha_{1k} - \mu \alpha_{2k}) + O(k^{-2}), \\ \text{and} \end{aligned}$$

$$\begin{aligned} \text{var } \mu_k &= (\alpha_2)^{-2} \text{var}(\alpha_{1k} - \mu \alpha_{2k}) + O(k^{-2}), \\ \text{as } k \rightarrow \infty. \end{aligned}$$

PROOF: By a Taylor series linear expansion,

$$\begin{aligned}
 \mu_k - \mu &= \alpha_{1k}/\alpha_{2k} - \alpha_1/\alpha_2 \\
 &= (\alpha_{1k} - \alpha_1) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} \\
 &\quad + \frac{1}{2} \left[(\alpha_{1k} - \alpha_1)^2 \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial (\alpha_1)^2} + (\alpha_{2k} - \alpha_2)^2 \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial (\alpha_2)^2} + 2(\alpha_{1k} - \alpha_1)(\alpha_{2k} - \alpha_2) \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial \alpha_1 \partial \alpha_2} \right] \\
 &\quad + k^{-3/2} \xi_k \\
 &= (\alpha_{1k} - \alpha_1)(\alpha_2)^{-1} - (\alpha_{2k} - \alpha_2)\mu(\alpha_2)^{-1} \\
 &\quad + (\alpha_{2k} - \alpha_2)^2 \mu(\alpha_2)^{-2} - (\alpha_{1k} - \alpha_1)(\alpha_{2k} - \alpha_2)(\alpha_2)^{-2} \\
 &\quad + k^{-3/2} \xi_k \\
 &= \alpha_2^{-1}(\alpha_{1k} - \mu \alpha_{2k}) \left[1 - \alpha_2^{-1}(\alpha_{2k} - \alpha_2) \right] + k^{-3/2} \xi_k.
 \end{aligned}$$

Therefore,

$$(\mu_k - \mu)^2 = \alpha_2^{-2}(\alpha_{1k} - \mu \alpha_{2k})^2 \left[1 - 2\alpha_2^{-1}(\alpha_{2k} - \alpha_2) \right] + k^{-2} \xi_k,$$

which implies that

$$\begin{aligned}
 \text{MSE } \mu_k &= \alpha_2^{-2} \text{var}(\alpha_{1k} - \mu \alpha_{2k}) - 2\alpha_2^{-3} \\
 &\quad \text{cov} \{ (\alpha_{1k} - \mu \alpha_{2k})^2, (\alpha_{2k} - \alpha_2) \} + k^{-2} \xi_k. \quad (\text{A.1})
 \end{aligned}$$

Now we will show that the covariance term in (A.1) is asymptotically negligible. Since

$$k^{1/2}(\alpha_{1k} - \mu \alpha_{2k}) \stackrel{d}{=} N(0, \sigma_3^2) + k^{-1/2} \xi_k$$

for some constant σ_3^2 , then

$$k(\alpha_{1k} - \mu \alpha_{2k})^2 \stackrel{d}{=} \sigma_3^2 \chi_1^2 + k^{-1/2} \xi_k,$$

where χ_1^2 denotes a chi squared random variable with one degree of freedom. Furthermore,

$$k^{1/2}(\alpha_{2k} - \alpha_2) \stackrel{d}{=} N(0, \sigma_2^2) + k^{-1/2} \xi_k.$$

If the signs on α_{ik} are negated for $i = 1, 2$, then $k(\alpha_{1k} - \mu \alpha_{2k})^2$ does not change but $k^{1/2}(\alpha_{2k} - \alpha_2)$ is negated. Therefore, by symmetry,

$$\text{cov} \{ k(\alpha_{1k} - \mu \alpha_{2k})^2, k^{1/2}(\alpha_{2k} - \alpha_2) \} = O(k^{-1/2})$$

as $k \rightarrow \infty$. Hence,

$$\text{cov} \{ (\alpha_{1k} - \mu \alpha_{2k})^2, (\alpha_{2k} - \alpha_2) \} = O(k^2) \quad (\text{A.2})$$

as $k \rightarrow \infty$. Combining (A.1) and (A.2) the first part of the lemma follows. Since Lemma A.1 implies that

$$\text{bias } u_k = O(k^{-1})$$

as $k \rightarrow \infty$, then the second part of this lemma follows.

Condition C: Defining $\alpha_{Ti} = \lim_{n \rightarrow \infty} \hat{\alpha}_i$ given U_T , the estimator, $\hat{\alpha}_i$, of α_i satisfies the following, for $i = 1, 2$:

$$E(\hat{\alpha}_i | U_T) - \alpha_{Ti} = O(n^{-1});$$

$$\text{Given } U_T, \hat{\alpha}_i - \alpha_{Ti} = O_p(n^{-1/2});$$

and

$$E(|\hat{\alpha}_i - \alpha_{Ti}|^3 | U_T) = O(n^{-3/2})$$

as $k \rightarrow \infty$.

Condition D: Given U_T

$$n^{1/2} \left[\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} - \begin{pmatrix} \alpha_{T1} \\ \alpha_{T2} \end{pmatrix} \right] \stackrel{d}{=} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) + \begin{pmatrix} n^{-1/2} \xi_n \\ n^{-1/2} \xi_n \end{pmatrix}$$

for some positive definite matrix Σ , where $\alpha_{Ti} = \lim_{n \rightarrow \infty} \hat{\alpha}_i$ given U_T . Also,

$$E(|\hat{\alpha}_i - \alpha_{Ti}|^3 | U_T) = O(n^{-3/2})$$

as $n \rightarrow \infty$, for $i = 1, 2$.

THEOREM A.1 Under conditions C and D, we have that

$$\text{var } \hat{\mu} = \alpha_2^{-2} E \text{var}(\hat{\alpha}_1 - \mu_T \hat{\alpha}_2 | U_T) + O(n^{-2} + N^{-1})$$

as $n \rightarrow \infty$, where $\mu_T = \alpha_{T1}/\alpha_{T2}$.

PROOF: First we determine $E \text{var}(\hat{\mu} | U_T)$. Under Condition D we apply Lemma A.2 to obtain

$$\text{var}(\hat{\mu} | U_T) = \alpha_{T2}^{-2} \text{var}(\hat{\alpha}_1 - \mu_{U_T} \hat{\alpha}_2 | U_T) + n^{-2} \xi_n. \quad (\text{A.3})$$

Since

$$\alpha_{T2}^{-2} = (\alpha_2)^{-2} + N^{-1/2} \xi_n$$

and

$$\text{var}(\hat{\alpha}_1 - \mu_T \hat{\alpha}_2 | U_T) = n^{-1} \xi_n,$$

then (A.3) implies that

$$E \text{var}(\hat{\mu} | U_T) =$$

$$\alpha_2^{-2} E \text{var}(\hat{\alpha}_1 - \mu_T \hat{\alpha}_2 | U_T) + O(n^{-2} + n^{-1} N^{-1/2}) \quad (\text{A.4})$$

as $n \rightarrow \infty$. Now we determine $\text{var} E(\hat{\mu} | U_T)$. Condition C and Lemma A.1 imply that

$$E(\hat{\mu} | U_T) = \mu_T + n^{-1} \xi_n = \mu + (n^{-1} + N^{-1/2}) \xi_n.$$

Hence,

$$\text{var } E(\hat{\mu} | U_T) = O(n^{-2} + N^{-1}) \quad (\text{A.5})$$

as $n \rightarrow \infty$. Combining (A.4) with (A.5) the result follows.

A.1 The post-stratified estimator

Here we derive the equations related to the post-stratified estimator, $\hat{\mu}_{ps}$, where $\hat{\alpha}_{ps(1)}$ and $\hat{\alpha}_{ps(2)}$ satisfy Conditions C and D. Note that

$$E(\hat{\alpha}_{ps(i)} | U_T) = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} N_{Tgh}^{-1} \sum_{k \in U_{Tgh}} y_{ik}$$

for $i = 1, 2$, and we define $\mu_{T,ps}^* = E(\hat{\alpha}_{ps(1)} | U_T) / E(\hat{\alpha}_{ps(2)} | U_T)$. Recall the definitions of α_i^* and μ^* in (4.1) and (4.2).

Derivation of (4.3), the asymptotic variance of $\hat{\mu}_{ps}$:

Since

$$\begin{aligned} & \text{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps} \hat{\alpha}_{ps(2)} | U_T, n_{gh}) \\ &= N^{-2} \sum_{h=1}^H \sum_{g=1}^G \frac{N_{gh}^2 [1 - n_{gh}/N_{Tgh}]}{n_{gh}(N_{Tgh} - 1)} \sum_{k \in U_{Tgh}} \left[y_{1k} - \mu_{T,ps} y_{2k} - N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} (y_{1j} - \mu_{T,ps} y_{2j}) \right]^2 \end{aligned} \quad (\text{A.6})$$

then

$$\begin{aligned} & E \text{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps} \hat{\alpha}_{ps(2)} | U_T, n_{gh}) \\ &= N^{-2} \sum_{h=1}^H \sum_{g=1}^G \frac{N_{gh}^2 \left[\left(\sum_{j \in U_h} p_j \right) - n_h \right]}{n_h \left(\sum_{j \in U_{gh}} p_j \right) \left[\left(\sum_{j \in U_{gh}} p_j \right) - 1 \right]} \\ & \quad \sum_{k \in U_{gh}} p_k \left[y_{1k} - \mu^* y_{2k} - \frac{\sum_{j \in U_{gh}} p_j (y_{1j} - \mu^* y_{2j})}{\sum_{j \in U_{gh}} p_j} \right]^2 \\ & \quad + O(n^{-2} + n^{-1} N^{-1/2}) \end{aligned} \quad (\text{A.7})$$

as $n \rightarrow \infty$. Also, since

$$\begin{aligned} & E(\hat{\alpha}_{ps(1)} - \mu_{T,ps} \hat{\alpha}_{ps(2)} | U_T, n_{gh}) \\ &= N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} N_{Tgh}^{-1} \sum_{k \in U_{Tgh}} (y_{1k} - \mu_{T,ps} y_{2k}), \end{aligned} \quad (\text{A.8})$$

then

$$E[\text{var} E(\hat{\alpha}_{ps(1)} - \mu_{T,ps} \hat{\alpha}_{ps(2)} | U_T, n_{gh}) | U_T] = 0. \quad (\text{A.9})$$

Since Theorem A.1 and (A.7) imply that

$$\begin{aligned} \text{var } \hat{\mu}_{ps} &= (\alpha_2^*)^{-2} E \text{var}(\hat{\alpha}_{ps(1)} - \mu_{T,ps} \hat{\alpha}_{ps(2)} | U_T, n_{gh}) \\ & \quad + O(n^{-2} + N^{-1}) \end{aligned} \quad (\text{A.10})$$

as $n \rightarrow \infty$, then (A.9) implies (4.3).

Derivation of (3.1), the estimated variance of $\hat{\mu}_{ps}$:

In light of (A.6) we have the estimator

$$\begin{aligned} & \widehat{\text{var}} \left(n_{gh}^{-1} \sum_{k \in s_{gh}} \{y_{1k} - \mu_{T,ps} y_{2k}\} | U_T, n_{gh} \right) \\ &= \frac{1 - n_{gh}/N_{Tgh}}{n_{gh}(n_{gh} - 1)} \sum_{k \in s_{gh}} \left[y_{1k} - \mu_{T,ps} y_{2k} - n_{gh}^{-1} \sum_{j \in s_{gh}} (y_{1j} - \mu_{T,ps} y_{2j}) \right]^2. \end{aligned}$$

Using (A.10) the result follows.

Derivation of (3.2), the estimated bias of $\hat{\mu}_{ps}$:

Lemma A.1 implies that

$$\hat{\mu}_{ps} - \mu^* = O(n^{-1})$$

as $n \rightarrow \infty$. Since

$$E \hat{\alpha}_{ps(i)} = \alpha_i^* + O(N^{-1})$$

as $N \rightarrow \infty$ for $i = 1, 2$, the result follows.

A.2 The phone-weighted estimator

Here we derive the equations related to the phone-weighted estimator, $\hat{\mu}_w$, under Conditions C and D, where $\tilde{\alpha}_{w(1)}$ and $\tilde{\alpha}_{w(2)}$ satisfy Conditions C and D. Note that

$$E(\tilde{\alpha}_{w(i)} | U_T) = N^{-1} \sum_{k \in U_T} y_{ik} / p_k$$

for $i = 1, 2$, and we define $\mu_{T,w} = E(\tilde{\alpha}_{w(1)} | U_T) / E(\tilde{\alpha}_{w(2)} | U_T)$.

Derivation of (4.4), the asymptotic variance of $\hat{\mu}_{ps}$:

Since

$$\begin{aligned} & \text{var}(\tilde{\alpha}_{w(1)} - \mu_{T,w} \tilde{\alpha}_{w(2)} | U_T) \\ &= N^{-2} \sum_{h=1}^H \frac{(N_{Th} - n_h) N_{Th}}{n_h (N_{Th} - 1)} \sum_{k \in U_{Th}} \left[\frac{y_{1k} - \mu_{T,w} y_{2k}}{p_k} - N_{Th}^{-1} \sum_{j \in U_{Th}} \frac{y_{1j} - \mu_{T,w} y_{2j}}{p_j} \right]^2, \end{aligned}$$

then

$$\begin{aligned}
& E \text{var}(\tilde{\alpha}_{w(1)} - \mu_{T,w} \tilde{\alpha}_{w(2)} | U_T) \\
&= N^{-2} \sum_{h=1}^H \frac{\left[\left(\sum_{j \in U_h} p_j \right) - n_h \right] \left(\sum_{j \in U_h} p_j \right)}{n_h \left[\left(\sum_{j \in U_h} p_j \right) - 1 \right]} \quad (\text{A.11})
\end{aligned}$$

$$\begin{aligned}
& \sum_{j \in U_h} p_k \left[\frac{y_{1k} - \mu y_{2k}}{p_k} - \frac{\sum_{j \in U_h} (y_{1j} - \mu y_{2j})}{\sum_{j \in U_h} p_j} \right]^2 \\
&+ O(n^{-1} N^{-1/2}) \quad (\text{A.12})
\end{aligned}$$

as $n \rightarrow \infty$. Applying Theorem A.1 to (A.11) the result follows.

Derivation of (3.6), the estimated variance of $\hat{\mu}_{ps}$:

In light of Theorem A.1 a valid estimate of $\text{var} \hat{\mu}_w$ also estimates

$$(\alpha_2)^{-2} E \text{var}(\hat{\alpha}_{w(1)} - \mu_{T,w} \hat{\alpha}_{w(2)} | U_T),$$

which is equivalent to

$$(\alpha_2)^{-2} E \text{var} \left(N^{-1} \sum_{h=1}^H \frac{N_{Th}}{n_h} \sum_{k \in s_h} \frac{y_{1k} - \mu_{T,w} y_{2k}}{p_k} \middle| U_T \right).$$

The result follows.

A.3 The post-stratified phone-weighted estimator

Here we derive the equations related to the post-stratified estimator, $\hat{\mu}_{psw}$, under Conditions C and D, where $\hat{\alpha}_{psw(1)}$ and $\hat{\alpha}_{psw(2)}$ satisfy Conditions C and D. Lemma A.1 implies that

$$E(\hat{\alpha}_{psw(i)} | U_T) = N^{-1} \sum_{h=1}^H \sum_{g=1}^G N_{gh} \frac{\sum_{k \in U_{Tgh}} p_k^{-1} y_{ik}}{\sum_{k \in U_{Tgh}} p_k^{-1}} + n^{-1} \xi_n$$

for $i = 1, 2$, and we define $\mu_{T,psw} = E(\hat{\alpha}_{psw(1)} | U_T) / E(\hat{\alpha}_{psw(2)} | U_T)$.

Derivation of (4.5), the asymptotic variance of $\hat{\mu}_{psw}$:

Using Lemma A.2 it follows that

$$\begin{aligned}
& \text{var} \left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,psw} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} \middle| U_T, n_{gh} \right) \\
&= \frac{1 - n_{gh} / N_{Tgh}}{n_{gh} (N_{Tgh} - 1)} \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-2}
\end{aligned}$$

$$\begin{aligned}
& \sum_{k \in U_{Tgh}} p_k^{-2} \left[y_{1k} - \mu_{psw, U_T} y_{2k} - \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-1} \right. \\
& \left. \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} \frac{y_{1j} - \mu_{T,psw} y_{2j}}{p_j} \right) \right]^2 + n^{-2} \xi_n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& E \text{var} \left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,psw} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} \middle| U_T, n_{gh} \right) \middle| U_T \\
&= \frac{(N_{Th} - n_h)}{n_h N_{Tgh} (N_{Tgh} - 1)} \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-2} \\
& \sum_{k \in U_{Tgh}} p_k^{-2} \left[y_{1k} - \mu_{T,psw} y_{2k} - \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} \right)^{-1} \right. \\
& \left. \left(N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} \frac{y_{1j} - \mu_{T,psw} y_{2j}}{p_j} \right) \right]^2 + n^{-2} \xi_n. \quad (\text{A.13})
\end{aligned}$$

Since

$$E N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} p_j^{-1} = \left(\sum_{k \in U_{gh}} p_k \right)^{-1} N_{gh} + O(N^{-1})$$

as $N \rightarrow \infty$, then (A.13) implies the unconditional expectation

$$\begin{aligned}
& E \text{var} \left(\frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,psw} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} \middle| U_T, n_{gh} \right) \\
&= \frac{N_{gh}^{-2} \left(\sum_{j \in U_{gh}} p_j \right) \left[\left(\sum_{j \in U_{gh}} p_j \right) - n_h \right]}{n_h \left[\left(\sum_{j \in U_{gh}} p_j \right) - 1 \right]} \\
& \sum_{k \in U_{gh}} p_k^{-1} \left[y_{1k} - \mu y_{2k} - N_{gh}^{-1} \sum_{j \in U_{gh}} (y_{1j} - \mu y_{2j}) \right]^2 \\
&+ O(n^{-2} + N^{-1}) \quad (\text{A.14})
\end{aligned}$$

as $n \rightarrow \infty$. By Theorem A.1,

$$\begin{aligned}
& \text{var} \hat{\mu}_{psw} \\
&= \alpha_2^{-2} E \text{var}(\hat{\alpha}_{psw(1)} - \mu_{T,psw} \hat{\alpha}_{psw(2)} | U_T, n_{gh}, \forall g, h) \\
&+ \alpha_2^{-2} E \text{var}[E(\hat{\alpha}_{psw(1)} - \mu_{T,psw} \hat{\alpha}_{psw(2)} | \\
& U_T, n_{gh}, \forall g, h) | n_{gh}, \forall g, h] + O(n^{-2} + N^{-1}) \quad (\text{A.15})
\end{aligned}$$

as $n \rightarrow \infty$. Lemma A.1 implies that

$$\text{var} \left[E(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} \mid U_T, n_{gh}, \forall g, h) \mid n_{gh}, \forall g, h \right] = n^{-2} \xi_n. \quad (\text{A.16})$$

Since (4.1) implies that

$$\begin{aligned} \text{var}(\hat{\alpha}_{\text{psw}(1)} - \mu_{T,\text{psw}} \hat{\alpha}_{\text{psw}(2)} \mid U_T, n_{gh}) \\ = N^{-2} \sum_{h=1}^H N_h^2 \\ \text{var} \left(N_h^{-1} \sum_{g=1}^G N_{gh} \frac{\sum_{k \in s_{gh}} p_k^{-1} \{y_{1k} - \mu_{T,\text{psw}} y_{2k}\}}{\sum_{k \in s_{gh}} p_k^{-1}} \mid U_{T,n_{gh}} \right), \end{aligned}$$

then (A.14), (A.15), and (A.16) imply the result.

Derivation of (3.7), the estimated variance of $\hat{\mu}_{\text{psw}}$:

Observe that

$$\frac{N_{Tgh} \sum_{j \in U_h} p_j}{n_{gh} n_h} \approx \frac{N_{Tgh} N_{Th}}{n_{gh} n_h} \approx \left[\frac{N_{gh}}{\sum_{k \in s_{gh}} p_k^{-1}} \right]^2.$$

Noting (4.5) the result follows.

Derivation of (3.8), another estimated variance of $\hat{\mu}_{\text{psw}}$:

Since

$$\frac{N_{gh}}{\sum_{k \in s_{gh}} p_k^{-1}} \approx \frac{N_{Tgh}}{n_{gh}} \approx \frac{N_{Th}}{n_h} \approx \frac{N_h}{\sum_{k \in s_h} p_k^{-1}},$$

the result follows from (3.7).

REFERENCES

- BRICK, J.M., WAKSBERG, J. and KEETER, S. (1994). Evaluating the use of data on interruptions in telephone service for nontelephone households. *Proceedings of the Survey Research Methods Section*, American Statistical Association 19-28.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.
- KHURSHID, A., and SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- McCULLAGH, P., and NELDER, J.A. (1991). *Generalized Linear Models*. New York: Chapman and Hall.
- RAO, J.N.K. (1997). Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics*, 25, 1-21.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- STEEH, C.G., GROVES, R.M., COMMENT, R. and HANSMIRE, E. (1983). Report on the survey research center's surveys of consumer attitudes. *Incomplete Data in Sample Surveys*, (Ed. W.G. Madow, H. Nisselson and I. Olkin), Academic Press, New York, 1.
- THORNBERRY, JR., O.T., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Surveys*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg). New York: John Wiley & Sons, Inc., 25-49.

Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement

YVES TILLÉ¹

ABSTRACT

The post-stratified estimator sometimes has empty strata. To address this problem, we construct a post-stratified estimator with post-strata sizes set in the sample. The post-strata sizes are then random in the population. The next step is to construct a smoothed estimator by calculating a moving average of the post-stratified estimators. Using this technique it is possible to construct an exact theory of calibration on distribution. The estimator obtained is not only calibrated on distribution, it is linear and completely unbiased. We then compare the calibrated estimator with the regression estimator. Lastly, we propose an approximate variance estimator that we validate using simulations.

KEY WORDS: Unbiased estimation; Calibration on a distribution function; Conditional inclusion probabilities; Weighting.

1. INTRODUCTION

It is possible during a survey by sampling to identify the values of an auxiliary character for all population units. This information may be available when the units are selected in a database containing other variables of interest. The temptation is then to calibrate the results of a survey on this auxiliary information. The decision is made either to retain from this auxiliary variable only certain functions (moments, sizes) for the purpose of using a calibration method (see for example Deville and Särndal 1992 or Estevao, Hidioglou and Särndal 1995), or this variable can be divided into classes with the view to using a post-stratified estimator.

If the decision is to opt for the post-stratified estimator, deciding on the strata divisions can be delicate. Theoretically, the strata must be defined prior to the selection of the sample. Where should the post-strata boundaries be placed? What size should the post-strata be? This latter question is the most embarrassing because the main problem with post-stratification is the possibility of obtaining empty post-strata. This means that the post-strata have to be large enough so that the probability of obtaining an empty post-stratum is negligible. These problems are not limited to post-stratified estimators. Indeed, it is also possible to have no regression or calibrated estimators for some samples.

Our objective is to define a new method of using auxiliary information in the population. This method is based on the definition of post-strata for which the number of units is set in the sample and not in the population. In this way, it is possible to import into the estimator complex auxiliary information resulting from knowledge of all of the values taken by the auxiliary variable, while avoiding both the problem of defining post-strata borders and the problem of empty post-strata.

This article is organized as follows. In section 2, the notation is defined and in section 3, we describe the principle of rank conditioning, which is used to define the unbiased estimators in section 4. In section 5, the smoothed estimator is defined, and a specific case is examined in detail in section 6. Section 7 contains an application of the estimation of a distribution function. In section 8, this new estimator is compared with the regression estimator and the estimator for a simple design without replacement. Computation of variance is discussed in section 9. As a result of the impossibility of providing an exact solution, an approximation is provided in section 10, which is tested by simulations in section 11. Lastly, general conclusions are presented in section 12.

2. NOTATION

We assume a population composed of N observation units, with the labelling being denoted as $\{1, \dots, k, \dots, N\}$. In this population, we are interested in a character of interest $Y_k, k \in U$. The objective is to estimate the total $Y = \sum_{k \in U} Y_k$. We select a random sample S of fixed size n by means of a simple random design without replacement. We denote I_k the random indicator variable, which takes the value 1 if the unit k is in the sample and 0 if not. The inclusion probabilities first order are therefore defined by $\Pr(k \in S) = \pi_k = n/N, k \in U$, and the second order inclusion probabilities by $\Pr(k, l \in S) = \pi_{kl} = n(n-1)/(N(N-1)), k \neq l \in U$.

We will be interested in the class of linear estimators of Y , which is written as

$$\hat{Y}_w = \sum_{k \in S} w_k Y_k,$$

¹ Yves Tillé, Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 827, 2002 Neuchâtel, Suisse. E-mail: yves.tille@unine.ch

where the weights w_k may depend on the sample S and therefore be random.

One of the possibilities is to take $w_k = 1/\pi_k = n/N$, which gives the Horvitz-Thompson estimator,

$$\hat{Y}_{HT} = \sum_{k \in S} \frac{Y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} Y_k,$$

which is unbiased.

We will be focussing instead on the more general class of conditionally weighted estimators (Tillé 1998, 1999a) where the units are weighted by inverses of conditional inclusion probabilities. If Z is some statistic, then the conditionally weighted estimator

$$\hat{Y}_Z = \sum_{k \in S} \frac{Y_k}{E(I_k|Z)} \quad (1)$$

is strictly unbiased if and only if $E(I_k|Z) > 0$, for all $k \in U$. In effect,

$$E(\hat{Y}|Z) = \sum_{k \in U} \frac{E(I_k|Z)Y_k}{E(I_k|Z)} = Y.$$

Since the estimator is conditionally unbiased, it is also unconditionally unbiased. Depending on which statistic Z is used, estimator (1) generalizes the stratified estimator as well as (a close approximation) the regression estimator (see Tillé 1998).

3. CONDITIONING ON RANKS

Let us now assume that the N values $X_1, \dots, X_k, \dots, X_N$ of an auxiliary character x are known for N units of the population. First, we assume that all of the X_k take separate values (this hypothesis will be removed in section 5). The rank R_k of unit k is

$$R_k = \#\{l \in U | X_l \leq X_k\}.$$

Moreover, we denote r_j , $j = 1, \dots, n$, the ordered population ranks of the n selected units in the sample, thus $r_1 < r_2 < \dots < r_{n-1} < r_n$. The r_j are random variables with a negative hypergeometric distribution (see Tillé 1999b).

The statistic used to define the conditional probabilities of inclusion is a subset of $\{r_1, \dots, r_j, \dots, r_n\}$. First, we define

- an integer q such that $2 \leq q \leq n$, defining the period,
- an integer b such that $2 \leq b$, defining the border,
- an integer l such that $b \leq l \leq b+q-1$, defining the interval.

The quantities q , b , and l serve to define a subset of indices:

$$E_l = \{r_l, r_{l+q}, r_{l+2q}, \dots, r_{l+hq}, \dots, r_{l+Hq}\},$$

for $l = b, \dots, b+q-1$.

For example, if $n = 18$, $q = 4$, $b = 3$, then

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\},$$

$$E_4 = \{r_4, r_8, r_{12}, r_{16}\},$$

$$E_5 = \{r_5, r_9, r_{13}\},$$

$$E_6 = \{r_6, r_{10}, r_{14}\}.$$

The conditional inclusion probability is computed in relation to one of the E_l .

The value of H is defined in such a way that $l+Hq \leq n-b+1$ and thus H is the largest integer such that $H \leq (n-b-l+1)/q$. It is clear that H depends on l .

The next step is to compute the inclusion probabilities:

$$E(I_k|E_l) = \begin{cases} 1 & \text{if } k \in E_l \\ \frac{q-1}{r_{l+Hq} - r_{l+(h-1)q} - 1} & \text{if } r_{l+(h-1)q} < k < r_{l+Hq}, h = 1, \dots, H \\ \frac{l-1}{r_l - 1} & \text{if } k < r_l \\ \frac{n - (l+Hq)}{N - r_{l+Hq}} & \text{if } k > r_{l+Hq} \end{cases}$$

These inclusion probabilities are thus relatively uneven. However, they are all positive, including the borders. It is important to use a border $b \geq 2$ so that the first and the last post-stratum are not empty.

4. CLASS OF UNBIASED ESTIMATORS

Since $E(I_k|E_l) > 0$, we can construct an estimator that is unbiased and even conditionally unbiased with respect to E_l . By denoting $y_1, \dots, y_j, \dots, y_n$ the n values taken by the units in the sample ordered according to the R_k , we obtain

$$\begin{aligned}
\hat{Y}_l &= \sum_{k \in S} \frac{Y_k}{E(I_k|E_l)} \\
&= \frac{r_l - 1}{l - 1} \sum_{j=1}^{l-1} y_j + y_l \\
&\quad + \sum_{h=1}^H \left(\frac{r_{l+hq} - r_{l+(h-1)q} - 1}{q - 1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j} + y_{l+hq} \right) \\
&\quad + \frac{N - r_{l+Hq}}{n - (l + Hq)} \sum_{j=l+Hq+1}^n y_j \\
&= N_{0|l} \hat{y}_{0|l} + y_l + \sum_{h=1}^H (N_{h|l} \hat{y}_{h|l} + y_{l+hq}) + N_{H+1|l} \hat{y}_{H+1|l}
\end{aligned}$$

where

$$N_{0|l} = r_l - 1,$$

$$N_{h|l} = r_{l+hq} - r_{l+(h-1)q} - 1, h = 1, \dots, H,$$

$$N_{H+1|l} = N - r_{l+Hq},$$

$$\hat{y}_{0|l} = \frac{1}{l-1} \sum_{j=1}^{l-1} y_j,$$

$$\hat{y}_{h|l} = \frac{1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j}, h = 1, \dots, H,$$

and

$$\hat{y}_{H+1|l} = \frac{1}{n - (l + Hq)} \sum_{j=l+Hq+1}^n y_j.$$

This estimator is in reality a post-stratified estimator where the sizes of the post-strata are set in the sample. Since $E(I_k|E_l) > 0$, \hat{Y}_l is strictly unbiased unconditionally and conditionally to E_p , which is clearly not the case for the traditional post-stratified estimator, because the latter has a non-zero probability of having an empty post-stratum. By setting the size of the post-strata in the sample, creating empty post-strata becomes impossible. The corresponding size of the post-stratum in the population is a random variable $N_{h|l}$.

The estimator \hat{Y}_l has another interesting property. By using the definition of the $E(I_k|E_l)$, we can quite easily show that

$$\sum_{k \in S} \frac{1}{E(I_k|E_l)} = N.$$

The estimator is thus calibrated on the size of the population. This property, which is also held by the Horvitz-Thompson estimator in simple designs, is therefore not lost. Units where the ranks are in E_l are called pivot units, and are assigned a weight equal to 1, which makes the weights very unequal. A downside to \hat{Y}_l is the use of widely dispersed weights. This problem can be resolved by smoothing the estimators.

5. SMOOTHING ESTIMATORS

To resolve the problem of the dispersion of the weights, we compute a moving average for the estimators as follows:

$$\hat{Y}_c = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{Y}_l.$$

\hat{Y}_c retains all of the properties of the \hat{Y}_l . This means that it is unbiased, calibrated on N and linear and can therefore be written as

$$\hat{Y}_c = \sum_{j=1}^n w_j y_j,$$

where $w_j =$

$$\begin{cases}
\frac{1}{q} \sum_{l=b}^{b+q-1} \frac{r_l - 1}{l - 1}, & j < b, \\
\frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{m^-(j+l-b-q)} - 1}{j+l-b-m^-(j+l-b-q)-1} + 1 \right), & b \leq j < b+q-1, \\
\frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right), & b+q-1 \leq j \leq n-b+2-q, \\
\frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{m^+(j+l-b)} - r_{j+l-b-q} - 1}{m^+(j+l-b)-j+l-b-q-1} + 1 \right), & n-b+2-q < j \leq n-b+1, \\
\frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N+1-r_{n+1-l}-1}{n+1-(n+1-l)-1} = \\
\frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N-r_{n+1-l}}{l-1}, & n-b+1 < j, \\
m^-(x) = \begin{cases} 0 & \text{if } x < b \\ x & \text{if not} \end{cases}, \\
m^+(x) = \begin{cases} n+1 & \text{if } x > n-b+1 \\ x & \text{if not} \end{cases}
\end{cases} \quad (2)$$

$$r_0 = 0, \text{ and } r_{n+1} = N + 1.$$

Under the apparent complexity arising from the specific treatment of the borders, the weighting system is relatively simple. In the case where we are not too close to the borders, it takes the value

$$w_j = \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right) \\ = \frac{1}{q(q-1)} \sum_{a=0}^{q-1} (r_{j+a} - r_{j+a-q}).$$

If all of the values of the auxiliary variable are not distinct, we can assign the unit ranks with common values randomly. For example, if $X_1 = 2, X_2 = 5, X_3 = 5, X_4 = 7, X_5 = 8$, we select with a probability $1/2$, between, ranks $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5$, or $R_1 = 1, R_2 = 3, R_3 = 2, R_4 = 4, R_5 = 5$. We then compute the smoothed estimator for each permutation, and we calculate their mean. The advantage of this method is that it preserves an unbiased estimator. In effect, for each possible permutation, the estimator is unbiased. In practice, it is not necessary to compute estimators for all of the permutations. We can calculate the estimator for one permutation and then simply equalize the weights of the units having the same values for the variable x .

6. CASE WHERE $q = 2, b = 2$

When $q = 2$, and $b = 2$, we obtain after a few calculations

$$\hat{Y}_c = \frac{1}{2} \left\{ \sum_{j=3}^{n-2} y_j (r_{j+1} - r_{j-1}) \right. \\ + \frac{r_3 + 2r_2 - 3}{2} y_1 + \frac{r_3 + 1}{2} y_2 \\ + \frac{r_{n+1} - r_{n-2} + 1}{2} y_{n-1} + \left. \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n \right\} \\ = \frac{1}{2} \left\{ \sum_{j=1}^n y_j (r_{j+1} - r_{j-1}) \right. \\ + y_1 \frac{r_3 - 3}{2} + y_2 \frac{2r_1 + 1 - r_3}{2} \\ + y_{n-1} \frac{r_{n+1} + r_{n-2} + 1 - 2r_n}{2} + y_n \left. \frac{r_{n+1} - r_{n-2} - 3}{2} \right\},$$

where $r_0 = 0$ and $r_{n+1} = N + 1$. This brings us to an estimator proposed by Ren (2000, page 140) and obtained using a calibration argument. The way in which the borders are managed is the only slight difference.

Example 1: With a population of size $N = 20$. Let us assume that the values of the variable of interest are found in Table 1. We also assume that the sample of size $n = 7$ is composed of the units with ranks $\{3, 7, 8, 11, 12, 15, 17\}$. If we take $q = 2, l = 2, b = 2$ we obtain $E_2 = \{r_2, r_4, r_6\} =$

$\{7, 11, 15\}$. We can then calculate $E(I_k | E_2 = \{7, 11, 15\})$. The conditional inclusion probabilities are as follows:

$$E(I_3 | E_2 = \{7, 11, 15\}) = 1/6,$$

$$E(I_7 | E_2 = \{7, 11, 15\}) = 1,$$

$$E(I_8 | E_2 = \{7, 11, 15\}) = 1/3,$$

$$E(I_{11} | E_2 = \{7, 11, 15\}) = 1,$$

$$E(I_{12} | E_2 = \{7, 11, 15\}) = 1/3,$$

$$E(I_{15} | E_2 = \{7, 11, 15\}) = 1,$$

$$E(I_{17} | E_2 = \{7, 11, 15\}) = 1/5.$$

Table 1

Example of a Population of Size $N = 20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	7	12	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	14	8	17	9	16	12	19	4	5	11	18	10

The estimator

$$\hat{Y}_0 = \sum \frac{y_k}{E(I_k | E_2 = \{7, 11, 15\})}$$

is therefore unbiased and conditionally unbiased. Further, it is linear and

$$\sum_{k \in S} \frac{1}{E(I_k | E_2 = \{7, 11, 15\})} = N.$$

However, if we take $q = 2, l = 3, b = 2$, we obtain $E_3 = \{r_3, r_5\} = \{8, 12\}$. Using the same example, we then compute $E(I_k | E_3 = \{8, 12\})$, and we obtain

$$E(I_3 | E_3 = \{8, 12\}) = 2/7,$$

$$E(I_7 | E_3 = \{8, 12\}) = 2/7,$$

$$E(I_8 | E_3 = \{8, 12\}) = 1,$$

$$E(I_{11} | E_3 = \{8, 12\}) = 1/3,$$

$$E(I_{12} | E_3 = \{8, 12\}) = 1,$$

$$E(I_{15} | E_3 = \{8, 12\}) = 2/8 = 1/4,$$

$$E(I_{17} | E_3 = \{8, 12\}) = 2/8 = 1/4.$$

The estimator

$$\hat{Y}_1 = \sum \frac{y_k}{E(I_k | E_3 = \{8, 12\})}$$

is also unbiased and linear.

Lastly, we compute the mean of the two estimators:

$$\hat{Y}_c = \frac{\hat{Y}_0 + \hat{Y}_1}{2}.$$

The weights are obtained simply by calculating the mean of the weights of estimators \hat{Y}_0 and \hat{Y}_1 , and have the values

$$w_3 = (6 + 7/2)/2 = 19/4,$$

$$w_7 = (1 + 7/2)/2 = 9/4,$$

$$w_8 = (3 + 1)/2 = 2,$$

$$w_{11} = (1 + 3)/2 = 2,$$

$$w_{12} = (3 + 1)/2 = 2,$$

$$w_{15} = (1 + 4)/2 = 5/2,$$

$$w_{17} = (5 + 4)/2 = 9/2.$$

Estimator \hat{Y}_c is linear and strictly unbiased.

7. APPLICATION TO THE ESTIMATION OF THE DISTRIBUTION

There are several ways to appropriately use auxiliary information to estimate a distribution function. A description of these techniques can be found in Ren (2000) and in Wu and Sitter (2001). The method that we suggest also makes it possible to estimate the distribution. The distribution in the population is defined by

$$F_1(y) = \frac{1}{N} \sum_{k \in U} I\{y_k \leq y\},$$

and can be estimated by

$$\hat{F}_1(y) = \frac{\sum_{k \in S} w_k I\{y_k \leq y\}}{\sum_{k \in S} w_k},$$

where $I\{y \leq y_k\}$ is the indicator function, and the w_k are the weights allocated to the units k which have the value $1/\pi_k = N/n$ for the Horvitz-Thompson estimator, and which are given in (2) for the calibrated estimator.

Note that the two functions are discrete, but that there are far fewer jumps in S than in U . To offset the differences in the distributions between the sample and the population, we have smoothed the distribution functions by using, as Deville (1995) did, a linear interpolation of the centres of the risers, which involves defining $F_2(y)$ by linking the points

$$\frac{1}{2} \{F_1(y_k) - F_1(y_k - \varepsilon)\},$$

for $k \in U$, where ε is a strictly positive, arbitrarily small real number. We then define $\hat{F}_2(y)$ by linking the points

$$\frac{1}{2} \{\hat{F}_1(y_k) - \hat{F}_1(y_k - \varepsilon)\},$$

for the sample.

Example 2: A population of size $N = 1\,000$ was generated using independent log-normal variables that are equally distributed. A sample of size $n = 16$ was then selected and we set $h = 5$. Figure 1 gives $F_2(x)$ in the population.

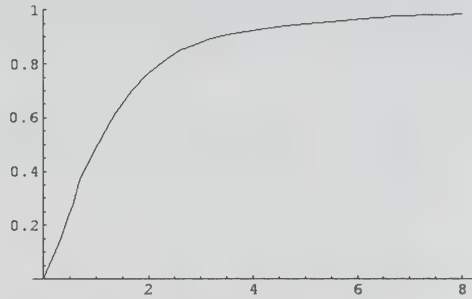


Figure 1. Population distribution function

Figure 2 shows $F_2(x)$ and the distribution estimated by the Horvitz-Thompson estimator. Lastly, Figure 3 shows $F_2(x)$ and the distribution estimated by the calibrated estimator.

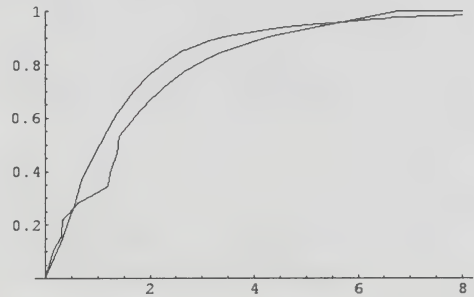


Figure 2. Population distribution function and Horvitz-Thompson distribution estimator

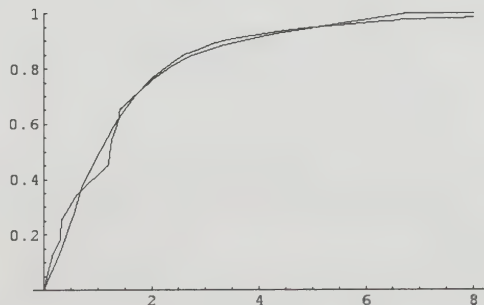


Figure 3. Population distribution function and calibrated distribution estimator

8. COMPARISON WITH THE REGRESSION ESTIMATOR

In order to compare the qualities of the proposed estimator, a series of simulations was conducted to compare the estimator calibrated on distribution with the Horvitz-Thompson estimator and the regression estimator. Three populations of size 1,000 were generated by means of the following models.

- *Model A Linear dependence*: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = X_k + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.616154$.
- *Model B Non-linear dependence 1*: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = (0.2 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.28975$.
- *Model C Non-linear dependence 2*: The population is generated using the model $X_k \sim N(0, 1)$ and $Y_k = (0.4 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0, 1)$. The coefficient of correlation obtained in the population is $\rho = 0.476158$.

In each population, 100,000 samples of size 100 were selected. Three weighting systems were computed for each sample.

1. the weights associated with the simple design $w_k = N/n$,
2. the weights of the distribution calibrated estimator given in (2) using the window $q = 10$ and border $b = 6$,
3. the weights of the regression estimator given by

$$w_k = \frac{N}{n} + (X - \hat{X}_{HT}) \frac{(X_k - \hat{X})}{\sum_{k \in S} (X_k - \hat{X})^2},$$

where X is the total of the X_k in the population, \hat{X}_{HT} is the Horvitz-Thompson estimator of X , and $\hat{X} = \hat{X}_{HT}/N$.

Using these weights, the estimator of the mean and of the nine deciles were calculated for each sample. We then estimate the variance of these estimators by means of the simulations.

The results are given in Tables 2, 3 and 4. The variances were brought to 1 for the simple design. For the linear model, the regression estimator is slightly preferable. However, in the non-linear case, the distribution calibrated estimator significantly increases the precision on the mean

and on the quantiles. This means that our proposed estimator is robust when there is a non-linear relationship between the auxiliary variable and the variable of interest.

Table 2
Model A: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.674422	0.632608
1st decile	0.905273	0.893876
2nd decile	0.815403	0.802082
3rd decile	0.842681	0.815071
4th decile	0.806749	0.768283
5th decile	0.783731	0.740765
6th decile	0.818051	0.782549
7th decile	0.794411	0.773794
8th decile	0.857114	0.844874
9th decile	0.884424	0.884032

Table 3
Model B: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.429689	0.953025
1st decile	0.913598	0.958656
2nd decile	0.919394	1.009270
3rd decile	0.829860	0.987950
4th decile	0.792094	0.989114
5th decile	0.703908	0.992023
6th decile	0.622705	1.009830
7th decile	0.550028	0.981249
8th decile	0.443828	1.010340
9th decile	0.549615	1.029120

Table 4
Model C: Estimator Variance
(Reference: Horvitz-Thompson=1)

Parameter	Distribution calibration	Regression estim.
Mean	0.30768	0.808114
1st decile	0.95560	0.983582
2nd decile	0.85920	0.970913
3rd decile	0.73854	0.930401
4th decile	0.65728	0.950651
5th decile	0.60500	0.956807
6th decile	0.52139	0.930514
7th decile	0.45709	0.907537
8th decile	0.40752	0.903593
9th decile	0.39820	0.860050

9. VARIANCE AND ESTIMATION OF VARIANCE

To compute the variance of \hat{Y}_c , we begin by computing the variance of \hat{Y}_l . Since \hat{Y}_l is unbiased conditionally to E_l , we have

$$V(\hat{Y}_l) = E V(\hat{Y}_l | E_l).$$

As with each of the post-strata, conditionally to E_l the design is a fixed-size simple sampling without replacement, we have

$$\begin{aligned} V(\hat{Y}_l | E_l) &= \sum_{h=0}^{H+1} N_{h|l}^2 V(\hat{y}_{h|l}) \\ &= \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{n|l}} \frac{S_{h|l}^2}{n_{h|l}}, \end{aligned} \quad (3)$$

where

$$n_{0|l} = l - 1,$$

$$n_{h|l} = q - 1, h = 1, \dots, H,$$

$$n_{H+1|l} = n - (l + Hq),$$

$$\bar{Y}_{0|l} = \frac{1}{N_{0|l}} \sum_{k=1}^{r_{l-1}} Y_{(k)},$$

$$\bar{Y}_{h|l} = \frac{1}{N_{h|l}} \sum_{k=r_{l-1} + (h-1)q+1}^{r_{l+hq-1}} Y_{(k)}, h = 1, \dots, H,$$

$$\bar{Y}_{H+1|l} = \frac{1}{N_{H+1|l}} \sum_{k=N-r_{l+Hq}+1}^N Y_{(k)},$$

$$S_{0|l}^2 = \frac{1}{N_{0|l} - 1} \sum_{k=1}^{r_{l-1}} (Y_{(k)} - \bar{Y}_{0|l})^2,$$

$$S_{h|l}^2 = \frac{1}{N_{h|l} - 1} \sum_{k=r_{l-1} + (h-1)q+1}^{r_{l+hq-1}} (Y_{(k)} - \bar{Y}_{h|l})^2, h = 1, \dots, H,$$

and

$$S_{H+1|l}^2 = \frac{1}{N_{H+1|l} - 1} \sum_{k=N-r_{l+Hq}+1}^N (Y_{(k)} - \bar{Y}_{H+1|l})^2,$$

where the $Y_{(k)}$ represent the values of Y_k sorted by increasing order of the X_k .

Note that it is very difficult to calculate the unconditional variance of \hat{Y}_l , that is, the expectation of $V(\hat{Y}_l | E_l)$. In effect, $N_{h|l}$ and $S_{h|l}^2$ are random. However, we can estimate $V(\hat{Y}_l | E_l)$ simply and obtain an unbiased estimator of the

conditional variance (and thus of the variance) by simply estimating (3), by

$$\hat{V}(\hat{Y}_l | E_l) = \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{n|l} n_{h|l}} s_{h|l}^2, \quad (4)$$

where

$$s_{0|l}^2 = \frac{1}{n_{0|l} - 1} \sum_{j=1}^{l-1} (y_j - \hat{y}_{0|l})^2,$$

$$s_{h|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=1}^{q-1} (y_{l+(h-1)q+j} - \hat{y}_{h|l})^2, h = 1, \dots, H,$$

and

$$s_{H+1|l}^2 = \frac{1}{n_{H+1|l} - 1} \sum_{j=l+Hq+1}^n (y_j - \hat{y}_{H+1|l})^2.$$

The estimator $\hat{V}(\hat{Y}_l | E_l)$ is not only unbiased for $V(\hat{Y}_l | E_l)$ but also for $V(\hat{Y}_l)$.

10. APPROXIMATIONS FOR COMPUTING THE VARIANCE

Unfortunately, computing the variance of \hat{Y}_c becomes more complex because of covariances. In effect, we have

$$V(\hat{Y}_c) = \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \text{Cov}(\hat{Y}_l, \hat{Y}_i).$$

When $l = i$, the problem is to estimate $V(\hat{Y}_l)$, which is done easily. When $l \neq i$, it is necessary to compute

$$\begin{aligned} \text{Cov}(\hat{Y}_l, \hat{Y}_i) &= E \text{Cov}(\hat{Y}_l, \hat{Y}_i | E_l) \\ &\quad + \text{Cov}(E(\hat{Y}_l | E_l), E(\hat{Y}_i | E_l)). \end{aligned}$$

Since $E(\hat{Y}_l | E_l) = Y$, we obtain

$$\text{Cov}(\hat{Y}_l, \hat{Y}_i) = E \text{Cov}(\hat{Y}_l, \hat{Y}_i | E_l)$$

$$= E E(\hat{Y}_l \hat{Y}_i | E_l) - Y^2.$$

Unfortunately, it does not appear possible to extricate the computation of $E(\hat{Y}_l \hat{Y}_i | E_l)$ and therefore we must find an approximation.

One way is to find a value that is greater than the variance. Since

$$\text{Cov}(\hat{Y}_l, \hat{Y}_i) \leq \sqrt{V(\hat{Y}_l) V(\hat{Y}_i)},$$

we have a greater value given by

$$\begin{aligned} V(\hat{Y}_c) &\leq \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \sqrt{V(\hat{Y}_l) V(\hat{Y}_i)} \\ &= \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{V(\hat{Y}_l)} \right)^2, \end{aligned}$$

which can be estimated by

$$\hat{V}_1(\hat{Y}_c) = \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{\hat{V}(\hat{Y}_l | E_l)} \right)^2,$$

which comes back to estimating the standard deviation of the means by the mean of the standard deviations.

Lastly, a second possibility involves using a residuals technique. Generally, when an estimator is corrected using a calibration technique, the variance is estimated by means of a residuals technique (see Deville and Särndal 1992 and Deville 1999 on this topic). When computing the variance of \hat{Y}_p , it is possible to use a residuals technique to obtain the exact variance. Consider the variable

$$v_k(l) = \begin{cases} \left(\frac{N^2(N-n)}{Nn(n-1)} \right)^{-1/2} \left(\frac{N_{h|l}^2(N_{h|l} - n_{h|l})}{N_{h|l}n_{h|l}(N_{h|l}-1)} \right)^{1/2} (Y_k - \bar{Y}_{h|l}) & \text{if } k = r_{l+(h-1)q+1}, \dots, r_{l+hq-1} \\ 0 & \text{if } k = r_{l+(h-1)q} \text{ or } k = r_{l+hq} \end{cases}$$

which can appear as a residual associated with the estimator \hat{Y}_l . The variable $v_k(l)$ inserted in the traditional expression of the fixed-size simple sampling design without replacement is exactly equal to the conditional variance \hat{Y}_l given in (3). In effect,

$$N^2 \frac{N-n}{nN} \frac{1}{N-1} \sum_{k \in U} \left(v_k - \frac{\sum_{k \in U} v_k}{N} \right)^2 = V(\hat{Y}_l | E_l).$$

This variable, however, depends on the $\bar{Y}_{h|l}$ which are unknown. We can estimate $v_k(l)$ by

$$\hat{v}_j(l) = \begin{cases} \left(\frac{N^2(N-n)}{Nn(n-1)} \right)^{-1/2} \left(\frac{N_{h|l}^2(N_{h|l} - n_{h|l})}{N_{h|l}n_{h|l}(n_{h|l}-1)} \right)^{1/2} (y_j - \hat{\bar{y}}_{h|l}) & \text{if } j = l + (h-1)q + 1, \dots, l + hq - 1 \\ 0 & \text{if } j = l + (h-1)q \text{ or } j = l + hq \end{cases}$$

If we insert $\hat{v}_k(l)$ in the estimator of the variance for the simple design without replacement, we obtain an unbiased estimator of the conditional variance, and therefore of the variance.

$$N^2 \frac{N-n}{nN} \frac{1}{n-1} \sum_{j=1}^n \left(\hat{v}_j - \frac{\sum_{j=1}^n \hat{v}_j}{n} \right)^2 = \hat{V}(\hat{Y}_l | E_l).$$

Deville (1999) shows that the variance of a sum of estimators can be determined by adding the residuals associated with these estimators, the residuals having been computed by linearization. To estimate the variance of \hat{Y}_c , we could therefore simply take the mean of the residuals $\hat{v}_k(l)$, which is written

$$\hat{v}_k = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{v}_k(l).$$

In this way, it would be possible to estimate the variance by

$$\hat{V}_2(\hat{Y}_c) = \frac{N^2(N-n)}{nN} \frac{1}{n-1} \sum_{k \in S} \left(\hat{v}_k - \frac{\sum_{k \in S} \hat{v}_k}{n} \right)^2.$$

These two variance estimators need to be tested by simulations.

11. SIMULATIONS FOR VARIANCE ESTIMATORS

The simulations shown in Table (5) are based on populations of size $N = 100$, that are generated by means of normal independent random variables. For each case studied, we give the value of q and the coefficient of correlation between the variable of interest Y_k and the rank R_k of the auxiliary variable X_k . The border b is defined by taking the integer of $q/2+1$. Since our purpose is to validate the variance estimator, we use 3,000 samples of size $n = 20$ for each simulation and we compare the variance estimated by the simulations of the calibrated estimator $V_{si}(\hat{Y}_c)$ with the expectations under the simulations of the two variance estimators denoted $E_{si}(\hat{V}_\alpha(\hat{Y}_c))$, $\alpha = 1, 2$. The last two columns of the tables show the relative bias defined by

$$RB_{si} \hat{V}_\alpha(\hat{Y}_c) = \frac{E_{si} \hat{V}_\alpha(\hat{Y}_c) - V_{si}(\hat{Y}_c)}{V_{si}(\hat{Y}_c)}, \quad \alpha = 1, 2.$$

The simulations show that the two proposed estimators overestimate the variance. The overestimation appears to diminish as q increases. The estimator $\hat{V}_2(\hat{Y}_c)$ definitely has the smallest bias. We will therefore prefer to use $\hat{V}_2(\hat{Y}_c)$.

Table 5
Simulation Results

Correlation: 0.802					
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.045	0.065	0.054	0.444	0.200
5	0.045	0.066	0.057	0.467	0.267
6	0.056	0.076	0.070	0.357	0.250
7	0.058	0.079	0.059	0.362	0.017
8	0.063	0.088	0.087	0.397	0.381
Correlation: 0.481					
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.048	0.066	0.059	0.375	0.229
5	0.045	0.060	0.054	0.333	0.200
6	0.044	0.056	0.051	0.273	0.159
7	0.044	0.054	0.051	0.227	0.159
8	0.045	0.052	0.048	0.156	0.067
Correlation: 0.111					
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$RB_{si}\hat{V}_1(\hat{Y}_c)$	$RB_{si}\hat{V}_2(\hat{Y}_c)$
4	0.281	0.471	0.363	0.676	0.292
5	0.297	0.420	0.356	0.414	0.199
6	0.279	0.363	0.316	0.301	0.133
7	0.267	0.342	0.324	0.281	0.213
8	0.282	0.327	0.281	0.160	-0.004

12. CONCLUSIONS

Our proposed estimator is one of the rare estimators that is both unbiased and linear, that uses auxiliary information and that is calibrated on the size of the population. It can be parameterized using the width of window q . This new estimator is robust compared with the regression estimator. It can take into account auxiliary information with a non-linear relationship with the variable of interest. Most simulations appear to show that the width of the window does not significantly impact the preciseness of the mean estimator. However, it also appears that a small window width is not penalizing, even if there is no dependence between the auxiliary variable and the variable of interest. The smaller q is, the tighter the calibration, and the variance estimator will then be significantly penalized because the degree of freedom is lost in each post-stratum. The choice of q must therefore reflect this problem.

There are many other methods that allow for the use of the information given by a distribution function (see Ren 2000) to improve an estimator. The results that we have presented are limited to simple sampling designs, but we

think they are important just as post-stratification is important as a specific case of calibration techniques. Post-stratification is one of the few examples where it is possible to show with accuracy that calibration corresponds to a conditional approach. Further, our approach can be seen as a calibration on a distribution function providing an unbiased estimator. A good general distribution calibration technique should therefore include in simple sampling designs the method we have presented.

ACKNOWLEDGEMENTS

We would like to thank Jean-Claude Deville and Anne-Catherine Favre, two referees and an associate editor for their constructive comments, which considerably improved this article.

REFERENCES

- DEVILLE, J.-C. (1995). *Estimation de la variance du coefficient de Gini mesuré par sondage*. INSEE Méthode, working paper, Methodology F9510.
- DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ESTEVAO, V., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principle for a generalized estimation system in Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- REN, R. (2000). *Estimation par calage sur la répartition*. Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.
- TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *International Statistical Review*, 66, 303-322.
- TILLÉ, Y. (1999a). Sur la détermination a posteriori des bornes des post-strates. In *Les Sondages* (Eds. G. Brossier and A.-M. Dussaix). Dunod, 202-208.
- TILLÉ, Y. (1999b). Estimation in surveys using conditional inclusion probabilities: complex design. *Survey Methodology*, 25, 57-66.
- WU, C., and SITTE, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289-307.

Variance Estimation for the Current Employment Survey

JUN SHAO and SHAIL BUTANI¹

ABSTRACT

Like most other surveys, nonresponse often occurs in the Current Employment Survey conducted monthly by the U.S. Bureau of Labor Statistics (BLS). In a given month, imputation using reported data from previous months generally provides more efficient survey estimators than ignoring nonrespondents and adjusting survey weights. However, imputation also has an effect on variance estimation: treating imputed values as reported data and applying a standard variance estimation method leads to negatively biased variance estimators. In this article we propose some variance estimators using the grouped balanced half sample method and re-imputation to take imputation into account. Some simulation results for the finite sample performance of the imputed survey estimators and their variance estimators are presented.

KEY WORDS: Balanced half samples; Non-negligible sampling fractions; Ratio imputation; Stratified sampling.

1. INTRODUCTION

The Current Employment Survey (CES), commonly known as the payroll survey, is conducted monthly by the U.S. Bureau of Labor Statistics (BLS). The data are obtained from establishments on a monthly basis by various automated methods including computer assisted telephone interviews, touchtone data entry, FAX, electronic data interchange, mail, *etc.* The main variables are the employment, production or non-supervisory workers and their working hours and earnings on nonagricultural establishment payrolls. Population employment counts are obtained once a year from Unemployment Insurance administrative records.

Nonresponse often occurs in the CES. In any particular month, imputation using reported data from previous months generally provides more efficient survey estimators than using reported data in the current month only and adjusting survey weights. This is particularly true in the CES because the nonresponse rate is about 60-80% and about 60% of the nonrespondents in a given month may become available one or several months later so that these data can be used as "reported data from previous months" (historical data) in a future month.

However, it is well known that treating imputed values as reported data and applying a standard variance estimation method leads to biased (often negatively biased) variance estimators. Valid variance estimators can be derived under some assumptions on sampling designs, imputation methods, and response mechanisms (and, sometimes, models that generate data).

The purposes of this article is to study variance estimation for the CES. After describing the sampling design and the imputation procedure currently used for the CES in section 2, we derive valid (asymptotically unbiased and consistent) variance estimators for imputed survey

estimators in section 3. To simplify the computation of variation estimators, we propose some approximations in section 4 and study their properties by simulation in section 5. Some conclusions are made in section 6. Although our study is motivated by the CES, we believe that our results are general and applicable to any survey that adopts a similar sampling design and a similar imputation method.

2. SAMPLING DESIGN AND IMPUTATION

The CES adopts the following stratified probability sampling design. Let P be a finite population containing a set of establishments $\{1, \dots, N\}$, which is stratified by the type of industry and by the size of the establishment. Within the h th stratum, a sample of size $n_h \geq 2$ is taken without replacement from N_h population units, using probability sampling independently across strata. The sampling fractions n_h/N_h are not necessarily negligible; for some strata with large establishment sizes, $n_h = N_h$. Let S denote the sample. For $i \in S$, at month $t = 0, 1, \dots, T$, values on the number of employees ($y_{t,i}^E$), the number of non-supervisory workers ($y_{t,i}^W$), the number of hours worked ($y_{t,i}^H$), and the weekly pay ($y_{t,i}^P$) are observed (if there is no nonresponse). Let $y_{t,i}$ denote any of $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$, or $y_{t,i}^P$. In CES, the main parameters of interest are population totals $Y_t = \sum_{i \in P} y_{t,i}$, $t = 1, \dots, T$. Since population totals can be obtained once a year from administrative records, we assume without loss of generality that Y_0 is known. If there is no nonresponse, Y_t is estimated by a ratio estimator

$$\hat{Y}_t = Y_0 \sum_{i \in S} w_i y_{t,i} / \sum_{i \in S} w_i y_{0,i}, \quad t = 1, \dots, T, \quad (1)$$

where w_i is the survey weight for the i th unit in the sample and the h th stratum.

¹ Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706; Shail Butani, Statistical Methods Division, The Bureau of Labor Statistics, Washington, D.C. 20212.

In our research, starting from month 1, nonrespondents are imputed using the imputation method proposed in Butani, Harter and Wolter (1997), as described below. Imputation is carried out within an imputation cell, which is the same as stratum or a union of strata. Imputed values in months 1, ..., $t-1$ are carried over to impute nonrespondents in month t , unless nonrespondents in months 1, ..., $t-1$ become respondents prior to month t .

1. The number of employees. If $y_{t,i}^E$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^E = \hat{\alpha}_t \tilde{y}_{t-1,i}^E,$$

where $\tilde{y}_{t-1,i}^E = y_{t-1,i}^E$ (reported value) if $y_{t-1,i}^E$ is available at month t and otherwise $\tilde{y}_{t-1,i}^E$ is an imputed value,

$$\hat{\alpha}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^E}{\sum_{j \in R_t} w_j y_{t-1,j}^E},$$

and R_t is the set of all reporting units for months t and $t-1$.

2. The number of non-supervisory workers. If $y_{t,i}^W$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^W = \tilde{y}_{t-1,i}^W \tilde{y}_{t,i}^E / \tilde{y}_{t-1,i}^E,$$

where $\tilde{y}_{t-1,i}^W$ is defined similarly to $\tilde{y}_{t-1,i}^E$.

3. The number of hours worked. If $y_{t,i}^H$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^H = \hat{\gamma}_t \tilde{y}_{t-1,i}^H \tilde{y}_{t,i}^W / \tilde{y}_{t-1,i}^W,$$

where $\tilde{y}_{t-1,i}^H$ is defined similarly to $\tilde{y}_{t-1,i}^E$ and

$$\hat{\gamma}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^H / \sum_{j \in R_t} w_j y_{t,j}^W}{\sum_{j \in R_t} w_j y_{t-1,j}^H / \sum_{j \in R_t} w_j y_{t-1,j}^W}.$$

4. The weekly pay. If $y_{t,i}^P$ is a nonrespondent, it is imputed by

$$\tilde{y}_{t,i}^P = \hat{\beta}_t \tilde{y}_{t-1,i}^P \tilde{y}_{t,i}^H / \tilde{y}_{t-1,i}^H,$$

where $\tilde{y}_{t-1,i}^P$ is defined similarly to $\tilde{y}_{t-1,i}^E$ and

$$\hat{\beta}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^P / \sum_{j \in R_t} w_j y_{t,j}^H}{\sum_{j \in R_t} w_j y_{t-1,j}^P / \sum_{j \in R_t} w_j y_{t-1,j}^H}.$$

Once nonrespondents are imputed, estimated monthly totals are calculated according to (1) by treating imputed values as reported data.

Assume that the population P is divided into K disjoint imputation cells P_1, \dots, P_K and for each k ,

$$y_{t,i} = \alpha_{t,k} y_{t-1,i} + \sqrt{y_{t-1,i}} e_{t,i},$$

$$E_m(y_{t,i}) = \mu_{t,k}, \quad E_m(e_{t,i}) = 0, \quad i \in P_k, \quad t = 1, 2, \dots,$$

$$V_m(y_{t,i}) = v_{t,k}, \quad V_m(e_{t,i}) = \sigma_k^2, \quad (2)$$

where $y_{t,i}$ denotes any of $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$, or $y_{t,i}^P$, E_m and V_m are the model (marginal) expectation and variance, respectively, $\alpha_{t,k}$ and σ_k^2 are unknown parameters, $e_{t,i}$'s are iid and the two processes $\{y_{t,i}\}$ and $\{e_{t,i}\}$ are independent. Within each P_k , it is assumed that the response indicator $a_{h,i}$ ($=1$ if $y_{t,i}$ is a respondent and $=0$ otherwise) and $y_{t,i}$ are independent, given $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, \dots, t$. Under this response mechanism, which is called unconfounded response mechanism (Lee, Rancourt and Särndal 1994), $a_{t,i}$ and $y_{t,i}$ are dependent, but through $y_{t-s,i}, a_{t-s,i}$, $s = 1, 2, \dots, t$. It is more general than the assumption that $(y_{1,i}, \dots, y_{t,i})$ and $(a_{1,i}, \dots, a_{t,i})$ are independent. Finally, response indicators from different units are assumed to be independent. Under these assumptions, the estimators \hat{Y}_t based on imputed data as described in the previous section are asymptotically unbiased with respect to the joint expectation under model (2) and sampling from the finite population.

In the CES, the imputation cells are unions of strata so that

$$\sum_{i \in S \cap P_k} w_i = M_k, \quad k = 1, \dots, K,$$

where M_k is the number of population units in the k th imputation cell P_k . Consequently, the \hat{Y}_t are conditionally unbiased with respect to the model expectation (given S), i.e.,

$$E_m(\hat{Y}_t - Y_t) = 0.$$

3. VARIANCE ESTIMATION

Let E_s and V_s be the sampling expectation and variance, respectively, and V be the overall variance. Then

$$\begin{aligned} V(\hat{Y}_t - Y_t) &= E_s[V_m(\hat{Y}_t - Y_t)] + V_s[E_m(\hat{Y}_t - Y_t)] \\ &= E_s[V_m(\hat{Y}_t - Y_t)], \end{aligned} \quad (3)$$

since $E_m(\hat{Y}_t - Y_t) = 0$. Furthermore, it is shown in the Appendix that

$$V_m(\hat{Y}_t - Y_t) = V_m(\hat{Y}_t) - V_m(Y_t). \quad (4)$$

Note that (4) is obvious in the case of no nonresponse.

Because of (3) the estimation of $V(\hat{Y}_t - Y_t)$ is the same as the estimation of $V_m(\hat{Y}_t - Y_t)$. Also, because of (4), we can first derive estimators v_{t1} and v_{t2} of $V_m(\hat{Y}_t)$ and $V_m(Y_t)$, respectively, and then take the difference $v_{t1} - v_{t2}$ as our variance estimator for \hat{Y}_t . Since $V_m(\hat{Y}_t)$ is a conditional variance, given S , we do not need to consider the sampling fractions n_h/N_h in the estimation of $V_m(\hat{Y}_t)$.

We first consider the estimation of $V_m(\hat{Y}_t)$. If an approximate formula of $V_m(\hat{Y}_t)$ can be derived, then we can directly estimate $V_m(\hat{Y}_t)$ by substitution. The explicit form of \hat{Y}_t , however, is very complex when t is not small so that the derivation of $V_m(\hat{Y}_t)$ is very difficult. Thus, in the CES we adopt a grouped half sample method that incorporates Rao and Shao's (1992) adjustment (or re-imputation) to take imputation into account. Specifically, sampled units in each stratum are randomly grouped into two groups. R half samples are created using a Hadamard matrix, where $H+1 \leq R \leq H+4$ and H is the number of strata. For the r th half sample and the i th sampled unit, define

$$w_i^{(r)} = \begin{cases} (1+0.5)w_i & \text{if the unit is in the } r\text{th} \\ & \text{half sample} \\ (1-0.5)w_i & \text{if the unit is not in the } r\text{th} \\ & \text{half sample,} \end{cases}$$

where w_i is the original survey weight. Let $\hat{Y}_t^{(r)}$ be the same as \hat{Y}_t except that the weights w_i are replaced by the $w_i^{(r)}$, including the weights used in imputation (i.e., $\hat{\alpha}_t, \hat{\gamma}_t$, and $\hat{\beta}_t$ are re-computed for every r , which is equivalent to Rao and Shao's adjustment). A grouped half sample variance estimator of $V_m(\hat{Y}_t)$ is

$$v_{t1} = \frac{4}{R} \sum_{r=1}^R \left(\hat{Y}_t^{(r)} - \frac{1}{R} \sum_{t=1}^R \hat{Y}_t^{(r)} \right)^2. \quad (5)$$

Note that the use of 0.5, instead of 1, in the construction of $w_i^{(r)}$ is based on Fay's method (Dippo, Fay and Morganstein, 1984; Judkins 1990; Rao and Shao 1999). Asymptotically, v_{t1} is unbiased and consistent for $V_m(\hat{Y}_t)$ (Shao, Chen, and Chen 1998; Rao and Shao 1999; Shao and Chen 1999).

We now consider the estimation of $V_m(Y_t)$. Under model (2),

$$V_m(Y_t) = \sum_k M_k v_{t,k},$$

which is of the order $O(N)$, where N is the size of the population P . Usually $V_m(\hat{Y}_t)$ is of the order $O(N^2/n)$, where $n = \sum_h n_h$ is the sample size. Hence $V_m(Y_t)/V_m(\hat{Y}_t)$ is of the order $O(n/N)$ and the estimation of $V_m(Y_t)$ is not necessary if n/N is negligible (although some sampling fractions n_h/N_h are not negligible).

In the CES, however, n/N is around 15% and is not negligible. Hence, the estimation of $V_m(Y_t)$ is necessary. An asymptotically unbiased and consistent estimator of $V_m(Y_t)$ is

$$v_{t2} = \sum_k M_k s_{k,t}^2, \quad (6)$$

where $s_{k,t}^2$ is the usual sample variance based on the respondents $y_{t,i}$ in the k th imputation cell.

4. APPROXIMATE VARIANCE ESTIMATORS

From section 3, a correct variance estimator for \hat{Y}_t is $v_{t1} - v_{t2}$, where v_{t1} and v_{t2} are given by (5) and (6), respectively. Although v_{t1} can be easily extended to the case where \hat{Y}_t is replaced by some nonlinear estimator such as \hat{Y}_t^P/\hat{Y}_t^H (the ratio of weekly pay over hour), the extension of v_{t2} involves the derivation of Taylor expansion for each separate nonlinear estimator. Thus, for the CES, it is desired to derive an approximate variance estimator that is not exactly correct but does not require the computation of v_{t2} .

Note that if n/N is negligible, then we can simply use v_{t1} as an estimator of $V(\hat{Y}_t - Y_t)$. In the CES, however, using v_{t1} leads to overestimation, since n/N is not negligible (see also the simulation results in section 5). Since this overestimation is caused by the sampling fraction, a possible way to fix the problem is to incorporate sampling fractions in the half sample method. When there is no nonresponse, sampling fractions can be incorporated into the half sample method by using formula (2) with $w_i^{(r)}$ replaced by

$$\tilde{w}_i^{(r)} = \begin{cases} (1+0.5\sqrt{1-n_h/N_h})w_i & \text{if the unit is in} \\ & \text{the } r\text{th half sample} \\ (1-0.5\sqrt{1-n_h/N_h})w_i & \text{if the unit is not} \\ & \text{in the } r\text{th half sample,} \end{cases} \quad (7)$$

when i is in stratum h .

Let \tilde{v}_{t1} be the variance estimator obtained using (5) but with $w_i^{(r)}$ replaced by $\tilde{w}_i^{(r)}$. If we use \tilde{v}_{t1} as an estimator of $V(\hat{Y}_t - Y_t)$, however, it has a negative bias, although it is better than the naive estimator that treats imputed values as observed data (see the simulation results in section 5).

While v_{t1} overestimates and \tilde{v}_{t1} underestimates the true variance $V(\hat{Y}_t - Y_t)$, a compromise is to replace the sampling fraction n_h/N_h in (7) by the "estimated sampling fraction" $r_{h,t}/N_h$, where $r_{h,t}$ is the number of respondents in stratum h at month t . Let \hat{v}_{t1} be the variance estimator obtained using (5) and (7) but with n_h/N_h in (7) replaced by $r_{h,t}/N_h$. Then

$$\tilde{v}_{t1} \leq \hat{v}_{t1} \leq v_{t1}.$$

All three variance estimators are asymptotically unbiased and are approximately equal when n/N is negligible. When n/N is not negligible, however, they are asymptotically biased.

To see the magnitude of the biases of \tilde{v}_{t1} , \hat{v}_{t1} , and v_{t1} , we consider the simplest case of no strata and $t = 1$. Let $y_i = y_{1,i}$, $x_i = x_{0,i}$ and

$$\hat{Y} = \sum a_i y_i + \sum (1 - a_i) \hat{R} x_i,$$

where $a_i = 1$ if y_i is a respondent and $a_i = 0$ otherwise, $\hat{R} = \sum a_i y_i / \sum a_i x_i$, and all summations are over $i \in S$. Let $\hat{U} = (\sum x_i / n) / (\sum a_i x_i / r)$, where r is the number of y -respondents. Then the correct variance estimator for \hat{Y} is $v_1 - v_2$ with

$$v_1 = \frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n}$$

and

$$v_2 = N \hat{U} s_d^2 + 2N \hat{U} \hat{R} s_{dx} + N \hat{R}^2 s_x^2,$$

where $s_d^2 = (r - 1)^{-1} \sum a_i (y_i - \hat{R} x_i)^2$, $s_{dx} = (r - 1)^{-1} \sum a_i x_i (y_i - \hat{R} x_i)$, and s_x^2 is the sample variance based on x_i 's. Also,

$$\begin{aligned} \tilde{v}_1 &= \left(1 - \frac{n}{N}\right) \left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n} \right) \\ &= v_1 - \frac{nN \hat{U}^2 s_d^2}{r} - 2N \hat{U} \hat{R} s_{dx} - N \hat{R}^2 s_x^2 \end{aligned}$$

and

$$\begin{aligned} \hat{v}_1 &= \left(1 - \frac{r}{N}\right) \left(\frac{N^2 \hat{U}^2 s_d^2}{r} + \frac{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}{n} \right) \\ &= v_1 - N \hat{U}^2 s_d^2 - \frac{2rN \hat{U} \hat{R} s_{dx} + rN \hat{R}^2 s_x^2}{n}. \end{aligned}$$

Since $v_1 - v_2$ is asymptotically unbiased, the bias of $v_{t1} = v_1$ is of the same order as v_2 and is always non-negative; the bias of $\tilde{v}_{t1} = \tilde{v}_1$ is of the same order as

$$N \hat{U} s_d^2 \left(1 - \frac{n}{r}\right) = -N \hat{U} s_d^2 \frac{\sum (1 - a_i) x_i}{\sum a_i x_i}$$

and is always non-positive; and the bias of $\hat{v}_{t1} = \hat{v}$ is of the same order as

$$N \hat{U} (1 - \hat{U}) s_d^2 + \left(1 - \frac{r}{n}\right) \left(2N \hat{U} \hat{R} s_{dx} + N \hat{R}^2 s_x^2\right) \quad (8)$$

The bias in (8) is non-negative if $s_{dx} \geq 0$ and $\hat{U} \approx 1$ (which is true if a_i is independent of x_i).

5. SOME SIMULATION RESULTS

To further study the biases of the variance estimators v_{t1} , \tilde{v}_{t1} and \hat{v}_{t1} , we conducted a simulation study using a CES dataset (from 1980's) of 149,044 units as the population P . Each unit $i \in P$ has a vector $y_i = (y_{t,i}^E, y_{t,i}^H, y_{t,i}^P, y_{t,i}^T, t = 0, 1, \dots, 7)$ and a vector r_i consisting of response indicators of the components of y_i , although all values of y_i are available (from administrative records). The sample S in the simulation was obtained by generating a stratified simple random sample $\{y_i\}$ of size 23,092 from P according to the sample allocations listed in Table 1. The response indicators of $\{y_i\}$ in the simulation were generated by drawing another (independent) stratified simple random sample $\{r_i\}$ from P . Thus, nonrespondents in the simulation were random and distributed according to the values of the r_i 's in the dataset P , but independent of the y_i 's.

After the sample data and nonrespondents were generated, nonrespondents were imputed as described in section 2. Estimated monthly totals \hat{Y}_t and monthly changes $\hat{Y}_t - \hat{Y}_{t-1}$ were calculated based on imputed data and their variance estimators, \tilde{v}_{t1} , \hat{v}_{t1} , v_{t1} , and $v_{t1} - v_{t2}$ were computed as described in sections 3 and 4. For comparison, the naive variance estimator v_{t0} , computed by treating imputed values as observed data, was also computed.

Based on 1,000 simulations, the relative biases (RB) and variances (Var) of the estimated totals \hat{Y}_t and changes $\hat{Y}_t - \hat{Y}_{t-1}$, the RB and coefficient of variations (CV) of the variance estimators for \hat{Y}_t and $\hat{Y}_t - \hat{Y}_{t-1}$, the coverage probability (CP) of the approximate 95% confidence intervals of the form

$$\text{the estimate} \pm 1.96 \sqrt{\text{the estimated variance}},$$

and the width (MW) of the confidence interval are given in Tables 2 through 5 respectively for 4 different variables. Estimated simulation standard errors are 2% for RB, CV, and MW, and 0.5% for CP.

Table 1
Sample Size by Stratum

SIC	SIZE	Stratum Size	Sample Size	Sampling Fraction	SIC	SIZE	Stratum Size	Sample Size	Sampling Fraction
10, 12-14	1	567	14	0.02439	50-51	1	3631	66	0.01812
	2	433	303	0.70000		2	3678	183	0.04987
	3	526	526	1.00000		3	4300	403	0.09375
	4	210	210	1.00000		4	1831	289	0.15789
	5	165	165	1.00000		5	833	320	0.38461
15-17	1	5055	129	0.02549	52-59	1	7084	149	0.02103
	2	4476	570	0.12731		2	5701	440	0.07724
	3	5281	1154	0.21854		3	8363	1037	0.12403
	4	2111	836	0.39583		4	4511	763	0.16915
	5	1005	1005	1.00000		5	4087	1002	0.24528
24-25, 32-29	1	3103	73	0.02349	60-62, 67	1	1384	17	0.01230
	2	3905	331	0.08475		2	971	38	0.03906
	3	6381	891	0.13966		3	1529	115	0.07500
	4	4273	1036	0.24242		4	981	67	0.06818
	5	4143	2127	0.51351		5	728	73	0.10000
20-23, 26-31	1	1754	40	0.02276	63-64	1	1364	15	0.01119
	2	1953	128	0.06564		2	652	20	0.03125
	3	3591	524	0.14599		3	754	87	0.11538
	4	3108	596	0.19167		4	435	48	0.11110
	5	3448	1041	0.30189		5	344	57	0.16667
40-49	1	1648	31	0.01902	7, 70-99	1	9641	230	0.02385
	2	1463	101	0.06918		2	6701	643	0.09602
	3	1988	221	0.11111		3	7833	1275	0.16275
	4	1171	211	0.18033		4	4839	1317	0.27215
	5	759	108	0.14286		5	4352	2067	0.47500

Table 2
Simulation Results for Employment

Estimation of total										Variance estimation for estimated total													
Month	Total*	RB	Var*	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	6.7E6	0.0	5.5E7	-37.0	47.6	85.3	7.7	-4.1	67.5	92.3	9.2	4.9	69.8	93.1	9.6	19.5	76.1	95.1	10.3	7.4	67.4	92.8	9.7
2	6.8E6	0.0	8.8E7	-34.3	28.8	86.9	9.6	-7.3	40.4	92.6	11.4	0.9	42.9	93.6	12.4	15.3	47.6	94.7	12.7	4.4	49.1	92.3	12.1
3	6.9E6	0.0	1.4E8	-26.1	30.4	88.2	12.9	-4.1	42.3	91.8	14.7	1.4	44.2	92.9	15.1	18.8	49.9	94.8	16.3	3.6	50.5	90.8	15.2
4	6.9E6	0.0	2.1E8	-22.5	32.9	89.3	16.1	-2.4	44.0	92.1	18.1	3.8	46.3	92.7	18.7	22.3	53.1	94.7	20.3	2.7	51.3	91.4	18.6
5	6.9E6	0.0	2.7E8	-21.9	35.0	88.3	18.4	-7.7	45.2	90.9	20.0	-1.1	47.9	92.0	20.7	16.2	55.6	94.4	22.4	-4.7	54.2	90.9	20.3
6	6.9E6	0.0	2.0E8	-8.8	40.5	91.7	17.1	-5.2	41.7	91.9	17.4	0.0	43.6	93.1	17.9	19.7	51.8	95.5	19.6	-3.1	52.5	90.5	17.6
7	6.9E6	0.0	1.5E8	-12.4	34.8	91.8	14.5	-8.6	36.1	92.5	14.8	-2.0	38.3	93.6	15.3	16.8	45.0	96.2	16.7	-6.6	42.4	92.7	15.0

Estimation of change										Variance estimation for estimated change													
Month	Change*	RB*	Var	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	8.0E4	-0.1	6.1E7	-43.0	25.4	84.9	7.5	-11.3	41.4	92.3	9.3	-4.5	43.9	93.7	9.7	9.4	48.7	95.6	10.3	8.6	51.7	93.5	10.3
3	9.7E4	-1.8	7.4E7	-35.0	31.7	85.0	8.7	-8.5	46.0	90.5	10.4	-3.2	47.7	91.0	10.7	11.7	53.1	93.4	11.5	-3.1	48.8	90.9	10.7
4	1.8E4	2.9	1.1E8	-31.8	42.3	87.4	11.0	-0.9	60.6	93.1	13.2	4.9	63.2	93.6	13.6	25.0	73.5	95.9	14.8	-2.5	47.7	89.9	13.1
5	4.4E4	3.4	1.1E8	-41.9	34.5	83.1	10.1	-10.8	57.3	91.4	12.5	-4.9	60.4	92.3	12.9	13.2	69.4	94.6	14.1	0.8	94.1	93.1	13.3
6	-1.1E4	9.3	1.1E8	-41.0	29.9	84.1	10.2	-12.6	42.0	91.1	12.4	-6.4	44.2	93.0	12.8	9.4	50.2	94.6	13.9	-4.1	53.9	93.0	13.0
7	1.6E3	3.2	1.2E8	-43.8	38.4	82.9	10.4	-15.9	57.5	89.6	12.7	-11.3	60.1	90.5	13.1	5.6	69.9	92.6	14.2	-0.2	75.5	90.0	13.8

Total: population total.

Change: population difference between the current month and the previous month.

Var: variance of the estimated total or change.

RB: relative bias = 100(bias/true value)%.

CV: coefficient of variation = 100 (standard error/true value)%.

CP: coverage probability of asymptotic confidence interval using estimated variance (in %).

MW: (mean width of asymptotic confidence interval)/10⁷.

*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 3
Simulation Results for Non-supervisory Workers

Month	Estimation of Total			Variance estimation															
	Total*	RB	Var*	v_{r0}				\tilde{v}_{r1}				\hat{v}_{r1}				v_{r1}			
				RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	5.4E6	-0.1	4.6E7	-33.3	49.7	80.9	7.0	-4.4	66.1	88.1	8.4	4.6	68.6	89.9	8.8	19.8	75.5	92.3	9.4
2	5.5E6	-0.1	7.6E7	-30.6	31.4	84.0	9.2	-7.4	41.1	89.4	10.6	0.9	43.7	91.0	11.1	15.8	48.7	93.8	11.9
3	5.6E6	-0.1	1.2E8	-23.6	31.2	85.6	12.1	-4.8	41.0	89.5	13.5	0.7	42.9	90.0	13.9	18.4	48.7	93.1	15.1
4	5.6E6	-0.1	1.9E8	-19.0	34.5	88.4	15.7	-2.4	43.8	91.7	17.2	3.8	46.3	91.9	17.8	22.5	53.2	94.1	19.3
5	5.7E6	-0.1	2.4E8	-18.9	36.8	87.8	17.6	-7.1	45.3	89.7	18.9	-0.4	48.2	90.7	19.6	17.2	56.0	93.0	21.2
6	5.7E6	0.0	1.8E8	-7.6	41.7	91.8	16.3	-4.7	42.8	92.4	16.6	0.6	44.8	92.7	17.0	20.6	53.1	95.4	18.6
7	5.7E6	0.0	1.4E8	-10.9	36.1	91.9	14.1	-7.7	37.2	92.2	14.4	1.0	39.4	93.6	15.0	18.3	46.3	95.9	16.3

Month	Estimation of Change			Variance estimation															
	Change*	RB	Var*	v_{r0}				\tilde{v}_{r1}				\hat{v}_{r1}				v_{r1}			
				RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	7.7E4	-0.8	5.1E7	-40.8	27.0	84.5	7.0	-12.9	41.2	91.5	8.4	-6.0	43.7	92.4	8.8	8.2	48.8	94.4	9.4
3	9.1E4	-1.4	6.2E7	-31.2	32.1	86.4	8.3	-8.7	42.8	91.2	9.5	-3.2	44.5	91.7	9.8	12.3	49.9	94.1	10.6
4	1.6E4	19.6	9.1E7	-27.2	44.0	87.1	10.3	-1.1	59.4	92.8	12.0	4.7	62.1	94.1	12.3	24.9	73.0	95.8	13.5
5	4.4E4	-0.4	9.5E7	-37.5	38.4	83.4	9.7	-10.0	58.6	90.8	11.7	-3.9	61.8	91.3	12.1	14.5	71.4	93.4	13.2
6	-1.0E4	-19.3	9.0E7	-37.0	32.4	83.4	9.5	-11.1	43.1	89.6	11.3	-4.7	45.5	90.4	11.7	11.7	51.8	92.4	12.7
7	7.9E2	48.7	1.0E8	-39.3	42.6	83.7	9.9	-14.5	59.7	89.2	11.7	-9.8	62.4	90.2	12.0	7.6	72.6	92.6	13.1

Total: population total.

Change: population difference between the current month and the previous month.

Var: variance of the estimated total or change.

RB: relative bias = 100(bias/true value)%.

CV: coefficient of variation = 100 (standard error/true value)%.

CP: coverage probability of asymptotic confidence interval using estimated variance (in %).

MW: (mean width of asymptotic confidence interval) / 10⁷.

*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 4
Simulation Results for Hours

Month	Estimation of Total			Variance estimation															
	Total*	RB	Var*	v_{r0}				\tilde{v}_{r1}				\hat{v}_{r1}				v_{r1}			
				RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
1	1.9E8	-0.1	5.8E10	-31.5	28.0	79.0	8.0	2.3	44.4	88.3	9.7	12.3	46.5	90.5	10.2	33.4	53.4	93.6	11.1
2	2.0E8	-0.1	1.2E11	-30.2	32.8	84.7	11.6	-7.7	40.4	90.6	13.3	0.1	42.8	91.7	13.9	19.7	49.4	94.3	15.2
3	2.0E8	-0.1	1.8E11	-23.3	30.0	86.3	14.9	-6.3	36.7	90.3	16.4	-1.0	38.1	91.2	16.9	19.6	43.8	94.6	18.6
4	2.0E8	0.0	3.2E11	-20.2	35.6	90.2	20.2	-0.5	47.1	93.4	22.6	5.6	49.7	93.3	23.3	27.9	59.8	95.3	25.6
5	2.1E8	0.0	4.4E11	-21.2	40.5	88.9	23.6	-7.9	52.3	90.7	25.5	-1.6	55.1	92.0	26.3	18.0	64.4	94.2	28.8
6	2.1E8	0.0	3.4E11	-10.4	46.3	92.1	22.1	-5.9	48.9	92.2	22.6	-1.0	50.7	93.0	23.2	20.8	59.9	94.7	25.6
7	2.1E8	0.0	2.3E11	-7.0	40.8	93.0	18.5	-2.2	42.8	93.2	19.0	4.2	44.7	94.1	19.6	27.2	53.2	95.8	21.6

Month	Estimation of Change			Variance estimation															
	Change*	RB	Var*	v_{r0}				\tilde{v}_{r1}				\hat{v}_{r1}				v_{r1}			
				RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW
2	5.0E6	0.1	8.8E10	-38.8	25.9	89.0	9.3	-9.7	35.1	92.4	11.3	-2.2	37.2	93.7	11.7	16.3	43.0	96.1	12.8
3	3.8E6	-1.0	1.1E11	-36.5	25.2	88.4	10.6	-12.6	34.5	91.9	12.4	-6.7	36.0	92.4	12.8	10.4	41.2	93.9	13.9
4	1.0E6	11.0	2.1E11	-31.2	45.6	87.3	15.2	-5.0	59.3	90.9	17.9	0.6	62.4	91.6	18.4	21.6	75.2	93.9	20.2
5	2.1E6	-0.5	2.2E11	-41.6	39.9	85.6	14.3	-14.3	63.9	91.1	17.4	-8.4	66.6	90.1	18.0	10.5	76.0	94.9	19.7
6	-7.7E5	-7.8	1.9E11	-40.1	35.1	82.5	13.5	-12.7	47.5	89.5	16.3	-6.5	50.3	90.7	16.9	12.7	60.1	94.1	18.5
7	2.5E5	-7.2	2.1E11	-39.0	48.4	82.9	14.3	-15.1	60.3	89.5	16.9	-10.6	62.4	90.3	17.3	8.0	72.3	94.0	19.0

Total: population total.

Change: population difference between the current month and the previous month.

Var: variance of the estimated total or change.

RB: relative bias = 100(bias/true value)%.

CV: coefficient of variation = 100 (standard error/true value)%.

CP: coverage probability of asymptotic confidence interval using estimated variance (in %).

MW: (mean width of asymptotic confidence interval) / 10¹⁰.

*: Scientific notation (for example, 6,700,000 is 6.7E6).

Table 5
Simulation Results for Weekly Pay

Estimation of Total												Variance estimation												
Month	Total*	RB	Var*	RB	v_{t0}				\tilde{v}_{t1}				\hat{v}_{t1}				v_{t1}				$v_{t1} - v_{t2}$			
					CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	
1	2.0E9	-0.1	9.5E12	-30.7	30.4	81.8	10.3	1.7	41.0	90.0	12.4	17.2	44.3	92.4	13.3	39.8	54.4	94.4	14.6	4.3	48.9	91.0	12.6	
2	2.1E9	-0.1	1.7E13	-27.2	27.8	84.3	14.1	-3.4	38.7	89.2	16.2	7.9	41.2	91.2	17.1	31.1	48.1	93.5	18.9	3.3	51.5	91.6	16.8	
3	2.1E9	-0.1	2.2E13	-14.3	34.7	85.6	17.4	1.1	42.2	88.1	18.9	8.0	43.9	89.5	19.5	34.9	51.4	93.5	21.8	2.6	50.4	91.4	19.0	
4	2.2E9	-0.1	3.7E13	-12.3	40.3	90.1	22.8	6.4	50.6	92.8	25.1	13.8	53.0	94.1	26.0	41.2	63.0	96.1	28.9	-0.9	84.5	92.8	24.2	
5	2.2E9	-0.1	5.0E13	-16.0	41.6	89.0	25.9	-1.5	51.8	91.4	28.1	5.9	54.8	92.0	29.1	29.3	64.6	94.3	32.2	-5.4	56.0	92.4	27.5	
6	2.2E9	-0.1	4.5E13	-9.4	44.1	92.0	25.5	-3.8	46.9	92.6	26.3	1.8	48.7	92.8	27.1	27.8	57.8	95.0	30.3	-0.4	54.1	94.2	26.8	
7	2.2E9	-0.1	3.5E13	-7.3	43.1	92.1	22.8	-0.7	48.3	92.8	23.6	6.8	50.0	93.9	24.5	31.9	57.0	96.4	27.2	-0.0	54.3	95.3	23.7	

Estimation of Change												Variance estimation												
Month	Change*	RB	Var*	RB	v_{t0}				\tilde{v}_{t1}				\hat{v}_{t1}				v_{t1}				$v_{t1} - v_{t2}$			
					CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	RB	CV	CP	MW	
2	6.4E7	-0.1	1.5E13	-37.6	25.7	85.4	12.2	-8.2	38.4	93.0	14.8	0.2	40.4	94.1	15.5	21.6	47.7	95.8	17.1	5.5	49.2	92.6	15.9	
3	3.5E7	-1.6	1.3E13	-31.7	27.9	87.7	11.9	-5.2	42.3	92.2	14.0	2.2	43.8	92.8	14.6	22.3	48.9	94.3	15.9	3.5	43.2	93.5	14.7	
4	2.1E7	6.6	2.4E13	-29.5	47.1	86.7	16.5	0.4	63.2	91.9	19.6	6.7	66.2	92.6	20.2	30.7	78.7	95.2	22.4	-4.3	96.9	90.6	19.2	
5	2.1E7	-0.4	2.4E13	-40.5	34.1	83.5	15.1	-9.2	55.7	90.5	18.7	-2.4	58.9	92.0	19.4	19.9	69.2	94.9	21.5	3.6	90.0	92.5	19.9	
6	1.4E7	2.0	2.3E13	-40.8	31.1	84.4	14.8	-13.5	46.0	91.4	17.8	-6.7	48.9	92.1	18.5	16.8	60.1	94.5	20.7	-4.4	53.0	91.5	18.8	
7	1.1E7	-0.1	2.7E13	-40.5	42.0	83.1	16.0	-13.9	56.5	89.2	19.3	-8.7	58.7	90.6	19.9	13.0	68.8	92.8	22.1	-3.7	69.5	90.8	20.4	

Total: population total.

Change: population difference between the current month and the previous month.

Var: variance of the estimated total or change.

RB: relative bias = 100(bias/true value)%.

CV: coefficient of variation = 100 (standard error/true value)%.

CP: coverage probability of asymptotic confidence interval using estimated variance (in %).

MW: (mean width of asymptotic confidence interval) / 10¹².

*: Scientific notation (for example, 6,700,000 is 6.7E6).

From Tables 2 through 5, the relative biases of estimators of monthly totals and changes are negligible for all variables. The following is a summary for the simulation results of variance estimators in terms of RB and CV.

1. As expected, the naive variance estimator v_{t0} has a large negative relative bias.
2. The asymptotically unbiased variance estimator $v_{t1} - v_{t2}$ performs well in general. Its relative bias is always under 10% in absolute value and is frequently under 5%.
3. The variance estimator v_{t1} has a large positive relative bias in all cases. This indicates that the v_{t2} term is not negligible in the CES in which the overall sampling fraction, n/N , is about 15%.
4. The variance estimator \tilde{v}_{t1} , which is the same as v_{t1} but with sampling fractions n_h/N_h incorporated (section 4), has a negative relative bias in general. Its negative bias may be large, especially in the estimation of the variance for monthly changes.
5. The variance estimator \hat{v}_{t1} , which is the same as \tilde{v}_{t1} but with sampling fractions n_h/N_h replaced by $r_{h,t}/N_h$, performs well in the simulation study, although it is not asymptotically unbiased (section 4). Its relative bias is large in a few cases, e.g., in variance estimation for total of weekly pay at months 1 and 4, in

variance estimation for total of hours at month 1, and in variance estimation for change of employment at month 7. In many cases, however, the performance of \hat{v}_{t1} is even better than the asymptotically unbiased estimator $v_{t1} - v_{t2}$.

The following is a summary for the simulation results of confidence intervals in terms of CP and MW.

1. The CP of the confidence interval based on the naive variance estimator v_{t0} is substantially lower than the nominal level 95% in most cases.
2. The CP of the confidence interval based on the asymptotically valid variance estimator, $v_{t1} - v_{t2}$, is between 90% and 93% in most cases. This is often the case for an asymptotically valid variance estimator, i.e., its relative bias is small but the CP of the related confidence interval is lower than the nominal level. One possible reason is that the convergence in distribution (asymptotic normality, which is the key for asymptotic confidence intervals) requires a larger sample size than the convergence of the second moment (in variance estimation).
3. In terms of CP, the confidence interval based on v_{t1} is the best. This might be because the overestimation in variance offsets the undercoverage in interval estimation. The mean width of the interval based on v_{t1} may

be substantially larger than those of other intervals, especially for weekly pay.

4. The CP of the confidence interval based on \hat{v}_{t1} , which is not asymptotically valid, is similar to that of the confidence interval based on $v_{t1} - v_{t2}$.

6. CONCLUSION AND DISCUSSION

For the survey estimators in the Current Employment Survey (CES) with imputed data, we propose an asymptotically unbiased and consistent estimator $v_{t1} - v_{t2}$ (section 3). Although v_{t1} can be easily computed using the grouped balanced half sample method, the computation of v_{t2} involves separate derivations for nonlinear estimators. Thus, several approximations, v_{t1} , \tilde{v}_{t1} , and \hat{v}_{t1} (section 4) are considered and compared with $v_{t1} - v_{t2}$ in a simulation study in which a CES dataset is used as population. Our result shows that v_{t1} and \tilde{v}_{t1} have large relative biases, due to the fact that the overall sampling fraction, 15%, is not negligible; the estimator \hat{v}_{t1} , which is the same as v_{t1} but incorporates an estimated sampling fraction (using the rate of response) in the balanced half sample method, performs fairly well. Thus, \hat{v}_{t1} is recommended to replace $v_{t1} - v_{t2}$ if the computation of v_{t2} is too complicated. Since the use of the "observed sampling fraction" $r_{h,t}/N_h$ does not reflect the fact that information is available about the nonrespondents from previous months, \hat{v}_{t1} may be improved using a more accurate estimated sampling fraction, for example, Rubin's (1987) "fraction of missing information".

Although our study is based on the CES, our results are applicable to any survey that adopts a similar sampling design and a similar imputation method. Furthermore, an extension to the case where model (2) involves $y_{t,i}, y_{t-1,i}, \dots, y_{t-s,i}$ with an integer $s \geq 2$ is straightforward, although the derivation of v_{t2} (for an asymptotically valid variance estimator) is more complicated.

ACKNOWLEDGEMENTS

The authors are grateful to an Associate Editor and two referees for their helpful comments and suggestions. The research of Jun Shao was partly supported by the NSF grant DMS-9803112 and DMS-01-02223 and the NSA grant MDA 904-99-1-0032.

APPENDIX: PROOF OF (4)

It suffices to show that

$$\text{Cov}_m(\hat{Y}_t, Y_t) = V_m(Y_t). \quad (9)$$

We show the case of a single imputation cell and $y_{t,i} = y_{t,i}^E$ (employment). The general case can be treated similarly.

We use mathematical induction. When $t = 1$,

$$\hat{Y}_t = \hat{\alpha}_1 Y_0.$$

By assumption (2),

$$\begin{aligned} \text{Cov}_m(\hat{Y}_t, Y_t) &= \alpha_1^2 V_m(Y_0) + \sigma^2 E_m(Y_0) \\ &= N(\alpha_1^2 v_0 + \sigma^2 \mu_0) \\ &= V_m(Y_1). \end{aligned}$$

Suppose now that (9) is true at time $t-1$. Let E_t , V_t and Cov_t be the expectation, variance and covariance conditional on $y_{j-1,i}, R_j, j=1, \dots, t$. Then

$$E_t(\hat{Y}_t) = \alpha_t \hat{Y}_{t-1}$$

and

$$\begin{aligned} \text{Cov}_t(\hat{Y}_t, Y_t) &= \text{Cov}_t(\hat{\alpha}_t \hat{Y}_{t-1}, Y_t) \\ &= \hat{Y}_{t-1} \text{Cov}_t(\hat{\alpha}_t, Y_t) \\ &= \sigma^2 \hat{Y}_{t-1}, \end{aligned}$$

where the last equality follows from assumption (2). By the induction assumption,

$$\text{Cov}_m(\hat{Y}_t, Y_{t-1}) = V_m(Y_{t-1}).$$

Then

$$\begin{aligned} \text{Cov}_m(\hat{Y}_t, Y_t) &= \text{Cov}_m[E_t(\hat{Y}_t), E_t(Y_t)] + E_m[\text{Cov}_t(\hat{Y}_t, Y_t)] \\ &= \alpha_t^2 \text{Cov}_m(\hat{Y}_{t-1}, Y_{t-1}) + \sigma^2 E_m(\hat{Y}_{t-1}) \\ &= \sigma_t^2 V_m(Y_{t-1}) + \sigma^2 E_m(Y_{t-1}) \\ &= V_m(Y_t). \end{aligned}$$

REFERENCES

- BUTANI, S., HARTER, R. and WOLTER, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 523-528.
- DIPPO, C.S., FAY, R.E. and MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the Section on Survey Research Methodology*, American Statistical Association. 489-494.

- LEE, H., RANCOURT, E. and SÄRNDAL C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- JUDKINS, D.R. (1990). Fay's method of variance estimation. *Journal of the Official Statistical*, 6, 223-239.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, 79, 811-822.
- RAO, J.N.K., and SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- SHAO, J., and CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhya, B*, Special Issue on Sample Surveys, 187-201.
- SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Implementing Rao-Shao Type Variance Estimation with Replicate Weights

MICHAEL P. COHEN¹

ABSTRACT

In estimating variances so as to account for imputation for item nonresponse, Rao and Shao (1992) originated an approach based on adjusted replication. Further developments (particularly the extension to balanced repeated replication from the jackknife replication of Rao and Shao) were made by Shao, Chen and Chen (1998). In this article we explore how these methods can be implemented using replicate weights.

KEY WORDS: Balanced Repeated Replication; Jackknife replication; Imputation; Item nonresponse; Weighted hot deck.

1. INTRODUCTION

Variance estimation by replication methods is facilitated by the use of replicate weights (Dippo, Fay and Morganstein 1984). In the past decade adjusted replication methods have been developed (Rao and Shao 1992; Shao, Chen and Chen 1998) that allow one to account for the variation due to imputation for item nonresponse in the estimation of variances. It is not, however, entirely obvious how these adjusted replication procedures can be implemented by means of replicate weights. This article explores how this can be done. The focus is on ways to prepare the dataset so that standard variance estimation software products that make use of replicate weights will work without modification. In the next to last section, however, some comments are made about whether modifying the software would help.

2. REPLICATION METHODS AND REPLICATE WEIGHTS

Wolter (1985) provides a comprehensive introduction to variance estimation for sample surveys. Chapters 3 and 4 cover the two replication methods pertinent to this article: the jackknife and balanced repeated replication. Shao and Tu (1995, chapter 6) is recommended for a more recent and advanced treatment. Variance estimation for surveys by replication continues to be an active area for research. Works that are even more recent include Brick and Morganstein (1996, 1997), Kott (2001), Rao and Shao (1996, 1999), Rust and Rao (1996), Shao (1996) and Valliant (1996).

The two replication methods work by creating subsets of the sample called *replicates*. The methods differ in the pattern by which replicates are formed. In balanced repeated replication (also called balanced half-sample replication), the replicates consist of roughly half the units

in the original sample; hence they are also called *half samples*. In jackknife replication (as applied to survey data), the replicates typically consist of the original sample except that a single primary sampling unit (PSU) or a small number of PSUs in the same stratum is deleted. For both methods, the replicates can be considered samples in their own right. Therefore if $\hat{\theta}$ is an estimate of some quantity θ based on the original sample, we can form an estimate $\hat{\theta}^{(r)}$ of θ based on replicate r . If there are R replicates, we estimate the sampling variance of $\hat{\theta}$, $\text{var}(\hat{\theta})$, by

$$\widehat{\text{var}}(\hat{\theta}) = C_{M,R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2 \quad (2.1)$$

where the constant $C_{M,R}$ depends solely on the replication method M and the number of replicates R .

In forming the estimate $\hat{\theta}$ of θ , use is made of the sample weights. For example, to estimate a population total for a particular item y , the estimate is the weighted sum of the values of y . Thus, if y_u and w_u are the values of y and the sample weight for sample unit u , then $\hat{\theta} = \sum_u w_u y_u$ where the sum is over all units in the sample. In addition to the sample weight w_u on the record for unit u , we can add replicate weights $w_u^{(r)}$, $r = 1$ to R , to the record on the file and calculate $\hat{\theta}^{(r)}$ in the same way as $\hat{\theta}$ except that $w_u^{(r)}$ replaces w_u for each sample unit u . Thus for the example in which $\hat{\theta}$ is the population total for y , $\hat{\theta}^{(r)} = \sum_u w_u^{(r)} y_u$. If unit u is not in replicate r , then $w_u^{(r)} = 0$. Some or all of the replicate weights for units that *are* in the replicate will be larger than their sample weights so that the units in the replicate continue to represent the entire population.

The use of replicate weights provided on the file to calculate the sampling variance estimates has advantages:

- Any statistics no matter how complicated that can be calculated for the whole sample can be calculated just as easily for each replicate. The sampling variance is then estimated by (2.1).

¹ Michael P. Cohen, Senior Mathematical Statistician, U.S. Bureau of Transportation Statistics, 400 Seventh Street SW, Washington, DC 20590 U.S.A.

- Adjustments for unit nonresponse and poststratification can (and should) be done individually for each replicate and incorporated in the replicate weights. This adjustment is usually done by an experienced sampling statistician and the adjusted replicate weights are put on the file so that the data analyst can use them without extra effort.
- Adjustments to the replicate weights put on the file can make use of auxiliary information not available to the data analyst, possibly for reasons of confidentiality. Even if not restricted, the auxiliary information may be difficult for the data analyst to obtain or use.
- General purpose software is available that employs replicate weights. Two software products that emphasize replication methods for surveys are WesVar from Westat, Inc. and VPLX from the U.S. Census Bureau. See the Web page

//www.fas.harvard.edu/~stats/survey-soft/survey-soft.html

for information on survey analysis software.

In this section we have ignored the complications that come from trying to capture the component of variance due to item imputation in the variance estimates. We begin to address these complications in the next section.

3. ADJUSTED REPLICATION METHODS

The works of Rao and Shao (1992) and Shao, Chen and Chen (1998) are key to this article. Shao and Chen (1999) and Shao and Steel (1999) also treat replication-based variance estimation for imputed survey data.

We begin by developing the notation, for the most part using that of Shao, Chen and Chen (1998). The population is divided into L strata with N_h clusters in the h th stratum. In the first stage of sampling in stratum h , $n_h \geq 2$ clusters are selected, the i th cluster being selected with probability p_{hi} , $i = 1, \dots, N_h$; $h = 1, \dots, L$. The clusters are selected without replacement and clusters in different strata are selected independently. The sampling fractions n_h/N_h are assumed to be small enough that no finite population correction is needed. Further stages of sampling may take place within each cluster, independently from cluster to cluster. There are N_{hi} ultimate population units in cluster i of stratum h . For population unit (h, i, j) , there is a variable y_{hij} of interest. Let S be the collection of all sample units and let $\{\tilde{y}_{hij}, (h, i, j) \in S\}$ be the imputed dataset: the \tilde{y}_{hij} are equal to y_{hij} when the item is observed and equal to the imputed value otherwise. The sample units are divided into *imputation classes* indexed by k and A_k is the index set of respondents for item y in imputation class k . We assume that the dataset contains identifiers ("flags") so that the nonrespondents can be identified.

In adjusted replication methods, \tilde{y}_{hij} in imputation class k is adjusted to

$$\tilde{y}_{hij}^{(r)} = \begin{cases} \tilde{y}_{hij} + E_{A_k}^{(r)}(\tilde{y}_{hij}) - E_{A_k}(\tilde{y}_{hij}) & \text{if } y_{hij} \text{ is imputed} \\ y_{hij} & \text{if } y_{hij} \text{ is observed,} \end{cases} \quad (3.1)$$

where E_{A_k} is the expectation with respect to the original imputation procedure within imputation class k and $E_{A_k}^{(r)}$ is the expectation with respect to the imputation procedure based only on data in the r th replicate within imputation class k . This formula is given explicitly in Shao, Chen and Chen (1998, page 822) for balanced repeated replication and a variety of imputation methods. It also applies to the development in Rao and Shao (1992) for jackknife replication and weighted hot deck imputation.

We shall adopt the notation that $(h^\circ i^\circ j^\circ)$ denotes a unit that did not respond to item y and $(h' i' j')$ denotes a unit that did respond to item y . We assume that

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h' i' j') \in A_k} a_{h' i' j'; h^\circ i^\circ j^\circ} y_{h' i' j'}$$

and

$$E_{A_k}^{(r)}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h' i' j') \in A_k} a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)} y_{h' i' j'}$$

where the $a_{h' i' j'; h^\circ i^\circ j^\circ}$ and $a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)}$ are constants not depending on the values of the $y_{h' i' j'}$ and $a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)} = 0$ for $(h' i' j')$ not in replicate r . The $a_{h' i' j'; h^\circ i^\circ j^\circ}$ and $a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)}$ may depend on auxiliary information available for all units in the sample. For the weighted hot deck of Rao and Shao (1992) and all of the imputation methods of Shao, Chen and Chen (1998), the expectations have this form.

3.1 Example: Ratio Imputation

This imputation method applies to situations in which there are auxiliary data $\{x_{hij}\}$ available for all sample units. Ratio imputation imputes a missing item $y_{h^\circ i^\circ j^\circ}$ by

$$x_{h^\circ i^\circ j^\circ} \sum_{(h' i' j') \in A_k} w_{h' i' j'} y_{h' i' j'} / \sum_{(h' i' j') \in A_k} w_{h' i' j'} x_{h' i' j'}$$

So

$$a_{h' i' j'; h^\circ i^\circ j^\circ} = x_{h^\circ i^\circ j^\circ} w_{h' i' j'} / \sum_{(h' i' j') \in A_k} w_{h' i' j'} x_{h' i' j'}$$

and

$$a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)} = x_{h^\circ i^\circ j^\circ} w_{h' i' j'}^{(r)} / \sum_{(h' i' j') \in A_k} w_{h' i' j'}^{(r)} x_{h' i' j'}$$

Notice that the $a_{h' i' j'; h^\circ i^\circ j^\circ}$ and $a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)}$ depend on the $\{x_{hij}\}$.

3.2 Example: Weighted Hot Deck Imputation

This imputation method imputes a missing item by a value randomly selected from the respondents to the same item with probability proportional to the weights of the respondents in the imputation class. See section 5 for further discussion of this method. Shao, Chen and Chen (1998, page 822) show that

$$E_{A_k}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h' i' j') \in A_k} w_{h' i' j'} y_{h' i' j'} / \sum_{(h' i' j') \in A_k} w_{h' i' j'}$$

and

$$E_{A_k}^{(r)}(\tilde{y}_{h^\circ i^\circ j^\circ}) = \sum_{(h' i' j') \in A_k} w_{h' i' j'}^{(r)} y_{h' i' j'} / \sum_{(h' i' j') \in A_k} w_{h' i' j'}^{(r)}.$$

Thus

$$a_{h' i' j'; h^\circ i^\circ j^\circ} = w_{h' i' j'} / \sum_{(h'' i'' j'') \in A_k} w_{h'' i'' j''}$$

and

$$a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)} = w_{h' i' j'}^{(r)} / \sum_{(h'' i'' j'') \in A_k} w_{h'' i'' j''}^{(r)}.$$

4. THE DATA FILE FOR VARIANCE ESTIMATION

For simplicity we assume that each record contains an identifier indicating to which imputation class the unit belongs. Often the imputation class is determined by several variables on the record. A record will look something like this:

$$ID \quad IC \quad w_{hij} \quad w_{hij}^{(1)} \quad \dots \quad w_{hij}^{(R)} \quad \tilde{y}_{hij} \quad IF_y \quad \tilde{z}_{hij} \quad IF_z$$

where ID is the identifier for the unit, IC is the identifier for the imputation class, w_{hij} is the (full sample) weight, $w_{hij}^{(1)} \dots w_{hij}^{(R)}$ are the replicate weights, \tilde{y}_{hij} is the value (possibly imputed) of the variable y under consideration, IF_y is the imputation "flag" that indicates whether \tilde{y}_{hij} is imputed, \tilde{z}_{hij} is the value (possibly imputed) of another variable z and IF_z is the imputation "flag" that indicates whether \tilde{z}_{hij} is imputed. There, of course, may be other variables on the files as well, for example an auxiliary variable x_{hij} available for all sample units.

We propose to add additional records, called extra records, to facilitate variance estimation. For each non-respondent ($h^\circ i^\circ j^\circ$) and respondent ($h' i' j'$) to item y in imputation class k , we create the record

$$ID \quad IC \quad 0 \quad \tilde{w}_{h^\circ i^\circ j^\circ; h' i' j'}^{(1)} \quad \dots \quad \tilde{w}_{h^\circ i^\circ j^\circ; h' i' j'}^{(R)} \quad y_{h' i' j'} \quad IF_y \quad 0 \quad IF_z$$

where $IC = k$, ID is the identifier of the unit ($h^\circ i^\circ j^\circ$) that did not respond to item y and

$$\tilde{w}_{h^\circ i^\circ j^\circ; h' i' j'}^{(r)} = (a_{h' i' j'; h^\circ i^\circ j^\circ}^{(r)} - a_{h' i' j'; h^\circ i^\circ j^\circ}) w_{h^\circ i^\circ j^\circ}^{(r)},$$

$$r = 1, \dots, R. \quad (4.1)$$

Note that the full sample weight is 0 on the extra records so these records do not affect the full sample estimates. The replicate estimates, though, agree with those defined by (3.1). Note also that the weights $\tilde{w}_{h^\circ i^\circ j^\circ; h' i' j'}^{(r)}$ may be negative.

Table 1
Numerical Illustration: Portion of Data File for Variance Estimation

ID	IC	w_{hij}	$w_{hij}^{(1)}$...	$w_{hij}^{(R)}$	\tilde{y}_{hij}	IF_y	\tilde{z}_{hij}	IF_z
001	1	10.1	20.2000	...	0.0000	5.4	1	1.2	1
002	1	20.3	40.6000	...	0.0000	5.1	0	1.3	0
003	1	18.4	36.8000	...	0.0000	5.2	0	1.3	0
004	1	11.1	0.0000	...	22.2000	5.1	1	1.2	0
005	1	16.3	0.0000	...	32.6000	5.1	1	1.4	0
006	1	15.4	0.0000	...	30.8000	5.4	0	1.4	0
001	1	0.0	3.0162	...	0.0000	5.1	2	0.0	3
001	1	0.0	2.7339	...	0.0000	5.2	2	0.0	3
001	1	0.0	-5.7501	...	0.0000	5.4	2	0.0	3
004	1	0.0	0.0000	...	-8.3301	5.1	2	0.0	3
004	1	0.0	0.0000	...	-7.5505	5.2	2	0.0	3
004	1	0.0	0.0000	...	15.8806	5.4	2	0.0	3
005	1	0.0	0.0000	...	-12.2325	5.1	2	0.0	3
005	1	0.0	0.0000	...	-11.0876	5.2	2	0.0	3
005	1	0.0	0.0000	...	23.3201	5.4	2	0.0	3
001	1	0.0	5.5645	...	0.0000	0.0	3	1.3	2
001	1	0.0	5.0436	...	0.0000	0.0	3	1.3	2
001	1	0.0	-2.7512	...	0.0000	0.0	3	1.2	2
001	1	0.0	-4.0400	...	0.0000	0.0	3	1.4	2
001	1	0.0	-3.8169	...	0.0000	0.0	3	1.4	2

Table 1 provides a numerical illustration. In the illustration, the nine records (rows of the table) with $IF_y = 2$ are the extra records for item y . The first six records are the original records for the six sample units that constitute imputation class $IC = 1$. (The records at the end with $IF_z = 2$ are the extra records for item z and will be discussed in the next paragraph. In these records, the imputation flag for y , IF_y , has been set to 3 to indicate that these are extra records for an item other than y .) There are three respondents ($IF_y = 0$) and three nonrespondents ($IF_y = 1$) to item y . The method of imputation is assumed to be weighted hot deck. Only the first and last replicate weights ($w_{hij}^{(1)}$ and $w_{hij}^{(R)}$) are presented, but these are consistent with replicate weights used for the balanced repeated replication method of variance estimation. We have $\sum w_{hij} \tilde{y}_{hij} = 476.650$, $\sum w_{hij}^{(1)} \tilde{y}_{hij} = 506.048$ and $\sum w_{hij}^{(R)} \tilde{y}_{hij} = 455.696$ where the sums are over all the

records. The reader may verify that this agrees with $\sum w_{hij} \tilde{y}_{hij} = 476.650$, $\sum w_{hij}^{(1)} \tilde{y}_{hij}^{(1)} = 506.048$ and $\sum w_{hij}^{(R)} \tilde{y}_{hij}^{(R)} = 455.696$ obtained using (3.1) where the sums are over the first six records only.

Let us now consider item z . The extra records for this item have the form

$$ID \quad IC \quad 0 \quad \tilde{w}_{h^{\circ}i^{\circ}j^{\circ};h'i'j'}^{(1)} \cdots \tilde{w}_{h^{\circ}i^{\circ}j^{\circ};h'i'j'}^{(R)} \quad 0 \quad IF_y \quad z_{h'i'j'} \quad IF_z$$

where $\tilde{w}_{h^{\circ}i^{\circ}j^{\circ};h'i'j'}^{(1)} \cdots \tilde{w}_{h^{\circ}i^{\circ}j^{\circ};h'i'j'}^{(R)}$ are computed by (4.1) but using the imputation method and response pattern for item z . The imputation method for z need not be the same as the imputation method for y but must be of the form discussed in section 3. In Table 1 the extra records for item z can be identified by having $IF_z = 2$. We have then $\sum w_{hij} \tilde{z}_{hij} = 120.130$, $\sum w_{hij}^{(1)} \tilde{z}_{hij} = 124.349$ and $\sum w_{hij}^{(R)} \tilde{z}_{hij} = 115.400$ where the sums are over all the records. This agrees with the sums obtained by (3.1).

Clearly the biggest disadvantage of this approach is the large number of extra records that have to be added to the file. This disadvantage is less severe when the imputation classes are small. (There are, however, many factors that go into determining the size of the imputation classes.) The advantages, on the other hand, include the following:

- The adjusted replicate estimates and variance estimates can be computed with any software designed to estimate variances by means of replicate weights.
- If there is another variable, say y' , with the same pattern of nonresponse and the very same method of imputation as y (that is, the same a and $a^{(r)}$ values), the computation of replicate estimates for y' can be accommodated without adding more records.
- One can make estimates over subdomains, even if they cut across imputation class boundaries.
- Suppose the method of imputation is the weighted hot deck. Then one estimates the variance of a derived variable, say $\log y$ where $y > 0$, by simply adding the derived variable to each record and computing replicate estimates based on it. (We shall have more to say about the weighted hot deck in the next section.)

The data analyst may choose to delete the extra records from a copy of the data file and use the reduced file to check for outliers, formulate hypotheses, etc. When it comes time to estimate variances, the extra records would be merged back in.

It should be pointed out that Rao and Shao (1992) proposed and evaluated their jackknife variance estimation method only for the estimation of totals (or means). One must be cautioned against the use of the approach for more complex statistics. In the same way, Shao, Chen and Chen (1998) proposed their balanced repeated replication variance estimation method for functions of totals and for quantiles so it should not be used for other statistics.

5. THE WEIGHTED HOT DECK

The use of the weighted hot deck method of imputation (e.g., Cox 1980) has a number of advantages so we devote a separate section to it. Rao and Shao (1992) concentrate on this imputation method and it is discussed also in Shao, Chen and Chen (1998). Under this method, a missing item is imputed by a value selected at random from the respondents to that item in the imputation class. The probability of selection is proportional to $w_{h'i'j'}$, the weight of the respondent. The respondents that have a positive probability of being selected are called *potential donors*; the non-respondent being imputed is the *recipient*. If there is more than one item on the file that will be imputed by the weighted hot deck, simplifications occur if one uses complete respondents (units who responded to all items) as potential donors and uses only one donor to impute all items requiring weighted hot deck imputation for a given recipient. (The donor is selected for each sample unit having *any* item for which there is item nonresponse.)

If each unit in an imputation class has the same chance of responding to an item, the weighted hot deck yields design consistent estimates of means, totals and sample quantiles. The imputations, moreover, will be "plausible" in the sense of looking like real data.

An advantageous feature of the weighted hot deck is that it is equivariant under one-to-one transformations. To explain equivariance, consider a derived variable d where $d = g(y)$ and g is a one-to-one function. Then, using the weighted hot deck, we impute item y of unit $(h^{\circ}i^{\circ}j^{\circ})$ that did not respond to the item by $\tilde{y}_{h^{\circ}i^{\circ}j^{\circ}}$ and use $g(\tilde{y}_{h^{\circ}i^{\circ}j^{\circ}})$ for d . This is equivalent to using the weighted hot deck to impute d by $\tilde{d}_{h^{\circ}i^{\circ}j^{\circ}}$ and using $g^{-1}(\tilde{d}_{h^{\circ}i^{\circ}j^{\circ}})$ for y . This feature of hot deck imputation is not shared by many other methods. For example, under *mean imputation* (in which the imputed value is the mean of the values for respondents in the imputation class), g would have to be linear for the equivariance property to hold. The pertinence of this to variance estimation by adjusted replicate methods is that when hot deck imputation is used, the data analyst can add $d = g(y)$ to the file and estimate variances for d as well as for y .

Suppose that the weighted hot deck is employed for several variables on a file and suppose that only complete respondents are used as potential donors. In this case, even if the patterns of nonresponse are different for the variables being imputed, the implementation of the adjusted replication by replicate weights described in the previous section can be carried out with the same set of extra replicate weights

$$\tilde{w}_{h^{\circ}i^{\circ}j^{\circ};h'i'j'}^{(r)} = (a_{h'i'j';h^{\circ}i^{\circ}j^{\circ}}^{(r)} - a_{h'i'j';h^{\circ}i^{\circ}j^{\circ}}) w_{h^{\circ}i^{\circ}j^{\circ}}^{(r)}$$

for each variable.

6. ALTERNATIVES

In this section we consider alternative methods including one that requires modifying the software.

6.1 First Alternative

One way to reduce the number of records is to have extra records of the form

$$ID' \quad IC \quad 0 \quad \tilde{w}_{h'i'j'}^{(1)} \quad \cdots \quad \tilde{w}_{h'i'j'}^{(R)} \quad y_{h'i'j'} \quad IF_y \quad 0 \quad IF_z$$

where ID' is the identifier of the *potential donor* unit ($h'i'j'$) that responded to item y , B_k is the index set of units not responding to item y in imputation class k and

$$\tilde{w}_{h'i'j'}^{(r)} = \sum_{(h^*i^*j^*) \in B_k} (a_{h'i'j'; h^*i^*j^*}^{(r)} - a_{h'i'j'; h^*i^*j^*}^{(r)}) w_{h^*i^*j^*}^{(r)},$$

$$r = 1, \dots, R.$$

Under this setup, for a given item there is only one extra record per potential donor. The chief disadvantage is that, because of the summation, estimates for subdomains that cut across imputation classes cannot be computed.

6.2 Second Alternative

Perhaps the most obvious implementation would be to add the $\tilde{y}_{hij}^{(r)}$ to the (hij) record and modify software to use the $\tilde{y}_{hij}^{(r)}$ rather than \tilde{y}_{hij} when computing replicate estimates. The chief drawbacks are (1) sophisticated reprogramming of software would be needed, (2) if multiple variables may require imputation, the number of fields needed expands greatly and (3) it is unclear how a data analyst would estimate the variance of a derived variable, say d , unless the $d_{hij}^{(r)}$ were put on the file in advance. The favorable features of this implementation are (1) no extra records are needed and (2) variance estimates for subdomains do not require additional work.

7. CONCLUDING REMARKS

The adjusted replication methods of Rao and Shao (1992) and Shao, Chen and Chen (1998) provide a way of computing variance estimates that account for imputation for item nonresponse. An important next step is the development of ways to facilitate the computation. This article explored implementations based on the use of replicate weights.

ACKNOWLEDGEMENTS

The idea for this article came from a question posed by Robert E. Fay at a Washington Statistical Society presentation. The author is also grateful to both referees, the associate editor and the editor for helpful comments. The author worked for the National Center for Education Statistics when the initial version of the article was written.

The views in this paper are those of the author and no official support by the U.S. Department of Education or the U.S. Department of Transportation is intended or should be inferred.

REFERENCES

- BRICK, J.M., and MORGANSTEIN, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.
- BRICK, J.M., and MORGANSTEIN, D. (1997). Computing sampling errors from clustered unequally weighted data using replication: WesVarPC. *Bulletin of the International Statistical Institute, Proceedings*, 1, 479-482.
- COX, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 721-726.
- DIPPO, C.S., FAY, R.E. and MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 489-494.
- KOTT, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RAO, J.N.K., and SHAO, J. (1996). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Society*, 91, 343-348.
- RAO, J.N.K., and SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.
- RUST, K., and RAO, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381-397.
- SHAO, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, 27, 203-254.
- SHAO, J., and CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhyā*, B, Special Issue on Sample Surveys, 187-201.
- SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Society*, 93, 819-831.
- SHAO, J., and STEEL, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Society*, 94, 254-265.
- SHAO, J., and TU, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- VALLIANT, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Variance Estimation for the General Regression Estimator

RICHARD VALLIANT¹

ABSTRACT

A variety of estimators of the variance of the general regression (GREG) estimator of a mean have been proposed in the sampling literature, mainly with the goal of estimating the design-based variance. Estimators can be easily constructed that, under certain conditions, are approximately unbiased for both the design-variance and the model-variance. Several dual-purpose estimators are studied here in single-stage sampling. These choices are robust estimators of a model-variance even if the model that motivates the GREG has an incorrect variance parameter.

A key feature of the robust estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. We also show that the delete-one jackknife implicitly includes the leverage adjustments and is a good choice from either the design-based or model-based perspective. In a set of simulations, these variance estimators have small bias and produce confidence intervals with near-nominal coverage rates for several sampling methods, sample sizes, and populations in single-stage sampling.

We also present simulation results for a skewed population where all variance estimators perform poorly. Samples that do not adequately represent the units with large values lead to estimated means that are too small, variance estimates that are too small, and confidence intervals that cover at far less than the nominal rate. These defects need to be avoided at the design stage by selecting samples that cover the extreme units well. However, in populations with inadequate design information this will not be feasible.

KEY WORDS: Confidence interval coverage; Hat matrix; Jackknife; Leverage; Model unbiased; Skewness.

1. INTRODUCTION

Robust variance estimation is a key consideration in the prediction approach to finite population sampling. Valliant, Dorfman, and Royall (2000) synthesize much of the model-based literature. In that approach, a working model is formulated that is used to construct a point estimator of a mean or total. Variance estimators are created that are robust in the sense of being approximately model-unbiased and consistent for the model-variance even when the variance specification in the working model is incorrect. In this paper, that approach is extended to the general regression estimator (GREG) to construct variance estimators that are approximately model-unbiased but are also approximately design-unbiased in single-stage sampling. A number of alternatives are compared including the jackknife and some variants of the jackknife. We will use a particular class of linear models along with Bernoulli or Poisson sampling as motivation for the variance estimators. However, some of these estimators can often be successfully applied in practice to single-stage designs where selections are not independent.

Associated with each unit in the population is a target variable Y_i and a p -vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ where $i = 1, \dots, N$. The population vector of totals of the auxiliaries is $\mathbf{T}_x = (T_{x1}, \dots, T_{xp})'$ where $T_{xk} = \sum_{i=1}^N x_{ki}$, $k = 1, \dots, p$. The general regression estimator, defined below, is motivated by a linear model in which the Y 's are independent random variables with

$$\begin{aligned} E_M(Y_i) &= \mathbf{x}_i' \boldsymbol{\beta} \\ \text{var}_M(Y_i) &= v_i. \end{aligned} \quad (1.1)$$

In most situations (1.1) is a "working" model that is likely to be incorrect to some degree.

Assume that a probability sample s is selected and that the selection probability of sample unit i is $P(\delta_i = 1) = \pi_i$ where δ_i is a 0-1 indicator for whether a unit is in the sample or not. We assume that the sample selection mechanism is ignorable. Roughly speaking, ignorability means that the joint distribution of the Y 's and the sample indicators, given the \mathbf{x} 's, can be factored into the product of the distribution for Y given \mathbf{x} and the distribution for the indicators given \mathbf{x} (see Sugden and Smith 1984 for a formal definition). In that case, model-based inference can proceed using the model and ignoring the selection mechanism.

The n -vector of targets for the sample units is $\mathbf{Y}_s = (Y_1, \dots, Y_n)'$, and the $n \times p$ matrix of auxiliaries for the sample units is $\mathbf{X}_s = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$. Define the diagonal matrix of selection probabilities as $\boldsymbol{\Pi}_s = \text{diag}(\pi_i)$, $i \in s$, and the diagonal matrix of model-variances as $\mathbf{V}_s = \text{diag}(v_i)$. The GREG estimator of the total, $T = \sum_{i=1}^N Y_i$, is then defined as the Horvitz-Thompson estimator or π -estimator, $\hat{T}_\pi = \sum_s Y_i / \pi_i$, plus an adjustment:

$$\hat{T}_G = \hat{T}_\pi + \hat{\mathbf{B}}' (\mathbf{T}_x - \hat{\mathbf{T}}_x) \quad (1.2)$$

where $\hat{\mathbf{B}} = \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s$ with $\mathbf{A}_{\pi s} = \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s$, and $\hat{\mathbf{T}}_x = \sum_s \mathbf{x}_i / \pi_i$. The GREG estimator can also be written as

¹ Richard Valliant, Westat, 1650 Research Boulevard, Rockville, MD 20850.

$$\hat{T}_G = \mathbf{g}_s' \Pi_s^{-1} \mathbf{Y}_s \quad (1.3)$$

with $\mathbf{g}_s = \mathbf{V}_s^{-1} \mathbf{X}_s \mathbf{A}_{\pi_s}^{-1} (\mathbf{T}_x - \hat{\mathbf{T}}_x) + \mathbf{1}_s$ and $\mathbf{1}_s$ being an n -vector of 1's. Expression (1.3) will be useful for subsequent calculations.

A variant of the GREG, referred to as a "cosmetic" estimator, was introduced by Särndal and Wright (1984) and amplified by Brewer (1995, 1999). A cosmetic estimator also has design-based and model-based interpretations. The variance estimators in this paper could also be adapted to cover cosmetic estimation.

Assuming that N is known, the GREG estimator of the mean is simply $\hat{Y}_G = \hat{T}_G / N$. We will concentrate on the analysis of \hat{Y}_G . (In some situations, particularly ones where multi-stage sampling is used, the population size is unknown and an estimate, \hat{N} , must be used in the denominator of \hat{Y}_G . The following analysis for the mean does not apply in that case.) Either quantitative or qualitative auxiliaries (or both) can be used in the GREG. If a qualitative variable like gender (male or female) is used, then two or more columns in \mathbf{X}_s will be linearly dependent, in which case a generalized inverse, denoted by $\mathbf{A}_{\pi_s}^{-}$, will be used in (1.2) and (1.3). Note that, although $\mathbf{A}_{\pi_s}^{-}$ is not unique, the GREG estimator \hat{Y}_G is invariant to the choice of generalized inverse. The proof is similar to Theorem 7.4.1 in Valliant *et al.* (2000).

The GREG estimator is model-unbiased under (1.1) and is approximately design-unbiased in large probability samples. Note that the model-unbiasedness requires only that $E_M(Y_i) = \mathbf{x}_i' \beta$; if the variance parameters in (1.1) are misspecified, the GREG will still be model-unbiased. On the other hand, if $E_M(Y_i)$ is incorrectly specified, the GREG is model-biased and the model mean squared error may contain an important bias-squared term. The estimation error of the GREG \hat{Y}_G is defined as

$$\hat{Y}_G - \bar{Y} = N^{-1} (\mathbf{a}_s' \mathbf{Y}_s - \mathbf{1}_r' \mathbf{Y}_r)$$

where $\bar{Y} = T/N$, $\mathbf{a}_s = \Pi_s^{-1} \mathbf{g}_s - \mathbf{1}_s$, \mathbf{Y}_r is the $(N - n)$ -vector of target variables for the nonsample units, and $\mathbf{1}_r$ is a vector of $N - n$ 1's. Next, suppose that the true model for Y_i is

$$E_M(Y_i) = \mathbf{x}_i' \beta$$

$$\text{var}_M(Y_i) = \psi_i, \quad (1.4)$$

i.e., the variance specification is different from (1.1) but $E_M(Y_i)$ is the same. Using the estimation error, the error-variance of \hat{Y}_G is then

$$\text{var}_M(\hat{Y}_G - \bar{Y}) = N^{-2} (\mathbf{a}_s' \Psi_s \mathbf{a}_s + \mathbf{1}_r' \Psi_r \mathbf{1}_r)$$

where the $n \times n$ covariance matrix for \mathbf{Y}_s is $\Psi_s = \text{diag}(\psi_i)$ and Ψ_r is the $(N - n) \times (N - n)$ covariance matrix for \mathbf{Y}_r . When the sample and population sizes are both large and the sampling fraction, $f = n/N$, is negligible, the error-variance is approximately

$$\text{var}_M(\hat{Y} - \bar{Y}) \approx N^{-2} \sum_{i \in s} a_i^2 \psi_i. \quad (1.5)$$

Note that this variance depends on the true variance parameters, ψ_i , and on the working model variance parameters, v_i , because v_i is part of a_i . Since \mathbf{a}_s is approximately the same as $\Pi_s^{-1} \mathbf{g}_s$ when selection probabilities are small, the error variance in that case is also approximately

$$\text{var}_M(\hat{Y}_G - \bar{Y}) \approx N^{-2} \sum_{i \in s} \frac{g_i^2}{\pi_i^2} \psi_i. \quad (1.6)$$

For model-based variance estimation, we will take either of the asymptotic forms in (1.5) or (1.6) as the target. However, when the sampling fraction is substantial, the term $\mathbf{1}_r' \Psi_r \mathbf{1}_r / N^2$ can be an important part of the error-variance and (1.5) or (1.6) may be poor approximations.

We will consider the design variance under two single-stage plans-Bernoulli and Poisson. In Poisson sampling, the indicators δ_i for whether a unit is in the sample or not are independent with $P(\delta_i = 1) = 1 - P(\delta_i = 0) = \pi_i$ (see Särndal, Swensson, and Wretman 1992, section 3.5, for a more detailed description). Bernoulli sampling is a special case of Poisson sampling in which each unit has the same inclusion probability. Under these two plans, the approximate design-variance of \hat{Y}_G is

$$\text{var}_\pi(\hat{Y}_G) \approx N^{-2} \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} E_i^2 \quad (1.7)$$

where $E_i = Y_i - \mathbf{x}_i' \mathbf{B}$ and $\mathbf{B} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}$ is the regression parameter estimator evaluated for the full finite population. Särndal (1996) recommends using the GREG in conjunction with sampling plans for which (1.7) is valid on the grounds that the variance (1.7) is simple and that the use of regression estimation can often more than compensate for the random sample sizes that are a consequence of such designs.

The Bernoulli and Poisson designs and the linear models (1.1) and (1.4) serve mainly as motivation for the variance estimators presented in sections 2 and 3. As noted by Yung and Rao (1996, page 24), it is common practice to use variance estimators that are appropriate to a design with independent selections or to a with-replacement design even when a sample has been selected without replacement. Likewise, variance estimators motivated by a linear model are often applied in cases where departures from the model are anticipated. This practical approach underlies the thinking in this paper and is illustrated in the simulation study reported in section 4.

2. VARIANCE ESTIMATORS

Our general goal in variance estimation will be to find estimators that are consistent and approximately unbiased under both a model and a design. Kott (1990) also

considered this problem. Note that the goal here is not the estimation of a combined (or anticipated) model-design variance,

$$E_M E_\pi \left\{ \left[(\hat{Y}_G - \bar{Y}) - E_M E_\pi (\hat{Y}_G - \bar{Y}) \right]^2 \right\}.$$

Rather we seek estimators that are useful for both $\text{var}_M(\hat{Y} - \bar{Y})$ and $\text{var}_\pi(\hat{Y})$. The arguments given here are largely heuristic ones used to motivate the forms of the variance estimators. Additional, formal conditions such as those found in Royall and Cumberland (1978) or Yung and Rao (2000) are needed for model-based and design-based consistency and approximate unbiasedness.

First, consider estimation of the approximate model-variance given in (1.5). In the following development, we assume that, as N and n become large,

- (i) $N \max(\pi_i) = O(n)$ and
- (ii) $\mathbf{A}_{\pi s}^i / N$ converges to a matrix of constants, \mathbf{A}_o .

A residual associated with sample unit i is $r_i = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \mathbf{x}_i' \hat{\mathbf{B}}$. The vector of predicted values for the sample units can be written as

$$\hat{\mathbf{Y}}_s = \mathbf{H}_s \mathbf{Y}_s \quad (2.1)$$

where $\mathbf{H}_s = \mathbf{X}_s \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \Pi_s^{-1}$. The predicted value for an individual unit is $\hat{Y}_i = \sum_{j \in s} h_{ij} Y_j$ where $h_{ij} = \mathbf{x}_i' \mathbf{A}_{\pi s}^{-1} \mathbf{x}_j / (v_j \pi_i)$ is the $(ij)^{\text{th}}$ element of \mathbf{H}_s . The matrix \mathbf{H}_s is the analog to the usual hat matrix (Belsley, Kuh and Welsch 1980) from standard regression analysis. The diagonal elements of the hat matrix are known as leverages and are a measure of the effect that a unit has on its own predicted value. Notice that the inverses of the selection probabilities are involved in (2.1), although these would have no role in purely model-based analysis.

The following lemma, which is a variation of some results in Lemma 5.3.1 of (Valliant *et al.* 2000), gives some properties of the leverages and the hat matrix.

Lemma 1. Assume that (i) and (ii) hold. For $\mathbf{H}_s = \mathbf{X}_s \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \Pi_s^{-1}$ the following properties hold for all $i \in s$:

- (a) $h_{ij} = O(n^{-1})$
- (b) \mathbf{H}_s is idempotent.
- (c) $0 \leq h_{ii} \leq 1$.

Proof: Since $h_{ij} = \mathbf{x}_i' \mathbf{A}_{\pi s}^{-1} \mathbf{x}_j / (v_j \pi_i)$, conditions (i) and (ii) imply that $h_{ij} = O(n^{-1})$. Part (b) follows from direct multiplication, using the definition of \mathbf{H}_s . To prove (c) note that $h_{ii} \geq 0$ since it is a quadratic form. Part (b) implies that $h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij} h_{ji}$ which can hold only if $h_{ii} \leq 1$.

Next, we write the residual as $r_i = Y_i(1 - h_{ii}) - \sum_{j \in s(i)} h_{ij} Y_j$ where $s(i)$ is the sample excluding unit i . Since $E_M(r_i) = 0$, we have $E_M(r_i^2) = \text{var}_M(r_i)$ and

$$E_M(r_i^2) = \Psi_i(1 - h_{ii})^2 + \sum_{j \in s(i)} h_{ij}^2 \Psi_j \quad (2.2)$$

under model (1.4). Using Lemma 1(a), we have $h_{ii} = o(1)$, $h_{ij}^2 = o(1)$, and consequently, $E_M(r_i^2) \approx \Psi_i$. Thus, in large samples, r_i^2 is an approximately unbiased estimator of the correct model-variance even though the variance specification in model (1.1) was incorrect. As a result, r_i^2 is a robust estimator of the model-variance for unit i regardless of the form of Ψ_i . A simple, robust estimator of the approximate model-variance (1.5) is then

$$v_{R1}(\hat{Y}_G) = N^{-2} \sum_s a_i^2 r_i^2 \quad (2.3)$$

which is a type of "sandwich" estimator (see, *e.g.*, White 1982). (Note that a formal argument that v_{R1} is robust would require conditions such that $n^{-1} E_M(v_{R1})$ and $n^{-1} N^{-2} \sum_s a_i^2 \Psi_i$ converge to the same quantity.) Another variance estimator, similar to v_{R1} if $\mathbf{a}_s \approx \Pi_s^{-1} \mathbf{g}_s$, is

$$v_{R2}(\hat{Y}_G) = N^{-2} \sum_s \frac{g_i^2}{\pi_i} r_i^2. \quad (2.4)$$

An estimator of the approximate design-variance in (1.7) is

$$v_\pi(\hat{Y}_G) = N^{-2} \sum_s \frac{1 - \pi_i}{\pi_i} r_i^2. \quad (2.5)$$

An alternative suggested by Särndal *et al.* (1989) as having better conditional properties is

$$v_{\text{ssw}}(\hat{Y}_G) = N^{-2} \sum_s \frac{1 - \pi_i}{\pi_i} g_i^2 r_i^2. \quad (2.6)$$

Another, similar estimator, used in the SUPERCARP software (Hidioglou, Fuller and Hickman 1980) and derived using Taylor series methods, is

$$v_T(\hat{Y}_G) = N^{-2} \frac{n}{n-1} \sum_s \left(\frac{g_i r_i}{\pi_i} - \frac{1}{n} \sum_s \frac{g_i r_i}{\pi_i} \right)^2. \quad (2.7)$$

As shown in the Appendix, the second term in parentheses in (2.7) converges in probability to zero under model (1.1). Thus, $v_T \approx v_{R2}$ in large samples.

When the selection probability of each unit is small, v_{ssw} will be similar to v_{R1} , v_{R2} , and v_π . All three will be approximately model-unbiased under (1.4) and approximately design-unbiased under Bernoulli and Poisson sampling. On the other hand, v_π is approximately design-unbiased but ignores the g_i coefficients and is biased under either model (1.1) or (1.4).

As a simple example, consider Bernoulli sampling with $\pi_i = n/N$ and the working model $E_M(Y_i) = x_i \beta$, $\text{var}_M(Y_i) = \sigma^2 x_i$. Then the GREG is the ratio estimator

$\hat{Y}_G = \bar{Y}_s \bar{x} / \bar{x}_s$ where \bar{x} is a finite population mean. The approximate model variance under the more general specification, $\text{var}_M(Y_i) = \psi_i$, is $(\bar{\Psi}_s / n) (\bar{x} / \bar{x}_s)^2$ where $\bar{\Psi}_s = \sum_{i=1}^N \psi_i / n$. The approximate design-variance is $(1-f)/(nN) \sum_{i=1}^N (Y_i - x_i \bar{Y} / \bar{x}_s)^2$ where \bar{Y} is a finite population mean. The estimator $v_{R2} = n^{-2} (\bar{x} / \bar{x}_s)^2 \sum_s (Y_i - x_i \bar{Y}_s / \bar{x}_s)^2$ is approximately unbiased for the model-variance and, because $\bar{x} / \bar{x}_s \rightarrow 1$ in large Bernoulli samples, v_{R2} is also approximately unbiased for the design-variance as long as f is small. In contrast, $v_\pi = n^{-2} (1-f) \sum_s (Y_i - x_i \bar{Y}_s / \bar{x}_s)^2$ is approximately design-unbiased but is model-unbiased only in balanced samples where $\bar{x} = \bar{x}_s$. Royall and Cumberland (1981) noted similar results for the ratio estimator in simple random sampling without replacement.

3. ALTERNATIVE VARIANCE ESTIMATORS USING ADJUSTED SQUARED RESIDUALS

The first alternative variance estimator we consider is the jackknife. The particular version to be studied is defined as

$$v_J = \frac{n-1}{n} \sum_{i=1}^n \left[\hat{Y}_{G(i)} - \hat{Y}_{G(\cdot)} \right]^2 \quad (3.1)$$

where $\hat{Y}_{G(i)}$ has the same form as the full sample estimator after omitting sample unit i . If the selection probability has the form $\pi_i = n p_i$, then (3.1) can be rewritten. Using the convention that the subscript (i) means that sample unit i has been omitted, we have

$$\begin{aligned} \hat{Y}_{G(i)} &= \hat{T}_{G(i)} / N, \hat{Y}_{G(\cdot)} = \sum_{i \in s} \hat{Y}_{G(i)} / n, \hat{T}_{G(i)} \\ &= \hat{T}_{\pi(i)} + \hat{\mathbf{B}}_{(i)}' (\mathbf{T}_x - \hat{\mathbf{T}}_{x(i)}), \end{aligned}$$

$$\begin{aligned} \hat{T}_{\pi(i)} &= n \sum_{i \in s(i)} Y_j / [\pi_j (n-1)], \hat{T}_{x(i)} \\ &= n \sum_{j \in s(i)} x_j / [\pi_j (n-1)], \text{ and} \end{aligned}$$

$$\hat{\mathbf{B}}_{(i)} = \mathbf{A}_{\pi s(i)}^{-1} \mathbf{X}_{s(i)}' \mathbf{V}_{s(i)}^{-1} \Pi_{s(i)}^{-1} \mathbf{Y}_{s(i)} \text{ with}$$

$$\mathbf{A}_{\pi s(i)} = \mathbf{X}_{s(i)}' \mathbf{V}_{s(i)}^{-1} \Pi_{s(i)}^{-1} \mathbf{X}_{s(i)}$$

Another more conservative, but asymptotically equivalent, version of the jackknife replaces $\hat{Y}_{G(i)}$ with the full sample estimator \hat{Y}_G . Design-based properties of the jackknife in (3.1) are usually studied in samples selected with replacement (see, e.g., Krewski and Rao 1981, Rao and Wu 1985, Yung and Rao 1996), but applied in practice to without-replacement designs. Note that for the linear estimator $\hat{Y}_\pi = N^{-1} \sum_{i \in s} Y_i / \pi_i$ in probability proportional to size without-replacement sampling, neither the jackknife, v_J , nor the approximations to v_J given later in this section, reduce to the usual Horvitz-Thompson or Yates-Grundy variance estimators.

With some effort we can write the jackknife in a form that involves the residuals and the leverages. The rewritten

form will make clear the relationship of the jackknife to the variance estimators in section 2. First, note the following equalities that are easily verified:

$$\hat{T}_{\pi(i)} = \frac{n}{n-1} \left(\hat{T}_{\pi(i)} - \frac{Y_i}{\pi_i} \right), \hat{T}_{x(i)} = \frac{n}{n-1} \left(\hat{\mathbf{T}}_{x(i)} - \frac{\mathbf{x}_i}{\pi_i} \right) \quad (3.2)$$

$$\begin{aligned} \mathbf{X}_{s(i)}' \mathbf{V}_{s(i)}^{-1} \Pi_{s(i)}^{-1} \mathbf{Y}_{s(i)} &= \mathbf{X}_s' \mathbf{V}_s^{-1} \Pi_s^{-1} \mathbf{Y}_s - \mathbf{x}_i Y_i / v_i \pi_i, \\ \mathbf{A}_{\pi s(i)} &= \mathbf{A}_{\pi s} - \mathbf{x}_i \mathbf{x}_i' / v_i \pi_i \end{aligned} \quad (3.3)$$

Using a standard formula for the inverse of the sum of two matrices, the slope estimator, omitting sample unit i , equals

$$\mathbf{B}_{(i)} = \hat{\mathbf{B}} + n^{-1} \sum_s \frac{\mathbf{A}_{\pi s}^{-1} \mathbf{x}_i r_i}{1 - h_{ii} v_i \pi_i}.$$

Details of this and the succeeding computations are sketched in the Appendix. After a considerable amount of algebra, we have

$$\hat{T}_{G(i)} - \hat{T}_{G(\cdot)} = -\frac{n}{n-1} (D_i - \bar{D}_s) + \frac{n}{n-1} F_i$$

where

$$D_i = \frac{g_i r_i}{\pi_i (1 - h_{ii})}$$

and F_i is defined in the Appendix. The jackknife in (3.1) is then equal to

$$\begin{aligned} v_J(\hat{Y}_G) &= N^{-2} \frac{n}{n-1} \times \\ &\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2 \sum_s F_i (D_i - \bar{D}_s) \right]. \end{aligned} \quad (3.4)$$

Expression (3.4) is an exact equality and could be used as a computational formula for the jackknife. This would sidestep the need to mechanically delete a unit, compute $\hat{Y}_{G(i)}$, and so on, through the entire sample.

In large samples the first term in brackets in (3.4) is dominant while the second and third are near zero under some reasonable conditions. Thus, in large samples the jackknife is approximated by $v_J(\hat{Y}_G) \approx N^{-2} \sum_s (D_i - \bar{D}_s)^2$, or, equivalently,

$$\begin{aligned} v_J(\hat{Y}_G) &\approx \frac{1}{N^2} \times \\ &\sum_s \left[\frac{g_i r_i}{\pi_i (1 - h_{ii})} \right]^2 - \frac{1}{N^2 n} \left[\sum_s \frac{g_i r_i}{\pi_i (1 - h_{ii})} \right]^2. \end{aligned} \quad (3.5)$$

As shown in the Appendix, the second term in (3.5) converges in probability to zero under model (1.1). Consequently, a further approximation to the jackknife is

$$v_J(\hat{Y}_G) \approx \frac{1}{N^2} \sum_s \left[\frac{g_i^2 r_i^2}{\pi_i^2 (1 - h_{ii})} \right]^2. \quad (3.6)$$

As (3.5) and (3.6) show, the jackknife implicitly incorporates the g_i^2 coefficients needed for estimating the model-variance. The right-hand side of (3.6) is itself an alternative estimator that we will denote by $v_J^*(\hat{Y}_G)$.

Yung and Rao (1996) also derived an approximation to the jackknife for the GREG in multistage sampling. For single-stage sampling, their approximation is equal to v_T , defined in (2.7), which is the same as (3.5) if the leverages are zero. Duchesne (2000) also presented a formula for the jackknife, which he denoted as \hat{V}_{JK2} , that involved sample leverages. The advantage of (3.4) is that it makes clear which parts of the jackknife are negligible in large samples. Duchesne also presented an estimator, denoted by \hat{V}_{JK2}^* , that is essentially the same as v_{R2} and is an approximation to the jackknife.

Expressions (3.5) and (3.6) explicitly show how the leverages affect the size of the jackknife. Weighted leverages, h_{ii} , that are not near zero will inflate v_J . Depending on the configuration of the x 's, this could be a substantial effect on some samples.

Since h_{ii} approaches zero with increasing sample size, v_J , v_{R2} , v_{SSW} , and v_T have the same asymptotic properties. In particular, the jackknife is approximately unbiased with respect to either the model or the design and is robust to misspecification of the variances in model (1.1). However, the factor $(1 - h_{ii})$ in (3.6) is less than or equal to 1 and will make the jackknife larger than the other variance estimators. This will typically result in confidence intervals based on the jackknife covering at a higher rate than ones using v_{R2} , v_{SSW} , or v_T .

Note, also, that if a without-replacement sample is used, and some first-order or second-order selection probabilities are not small, the choices, v_{R2} , v_D , v_J , and v_J^* will be over-estimates of either the design-variance or the model-variance. To account for non-negligible selection probabilities, we can make some simple adjustments. An adjusted version of $v_J^*(\hat{Y}_G)$, patterned after v_{SSW} , is

$$v_{JP}^*(\hat{Y}_G) = \frac{1}{N^2} \sum_s \frac{(1 - \pi_i) g_i^2 r_i^2}{\pi_i^2 (1 - h_{ii})^2}.$$

This expression is similar to \hat{V}_{JK3}^* of Duchesne (2000), although \hat{V}_{JK3}^* omits the leverages. Expression (3.6) also suggests another alternative that is closely related to an estimator of the error variance of the best linear unbiased predictor of the mean under model (1.1) (see, Valliant *et al.* 2000, chapter 5). This estimator is somewhat less conservative than (3.6), but still adjusts using the leverages:

$$v_D(\hat{Y}_G) = \frac{1}{N^2} \sum_s \frac{g_i^2 r_i^2}{\pi_i^2 (1 - h_{ii})}.$$

Because $h_{ii} = o(1)$, v_D is also approximately model and design-unbiased. A variant of this that may perform better when some selection probabilities are large is

$$v_{DP}(\hat{Y}_G) = \frac{1}{N^2} \sum_s \frac{(1 - \pi_i) g_i^2 r_i^2}{\pi_i^2 (1 - h_{ii})}.$$

4. SIMULATION RESULTS

To check the performance of the variance estimators, we conducted several simulation studies using three different populations. The first is the Hospitals population listed in Valliant *et al.* (2000, Appendix B). The second population is the Labor Force population described in Valliant (1993). The third is a modification of the Labor Force population. In all three populations, sampling is done without replacement, as described below. These sampling plans will test the notion that variance estimators motivated, in part, by with-replacement designs can still be useful when applied to without-replacement designs.

The Hospitals population has $N = 393$ and a single auxiliary value x , which is the number of inpatient beds in each hospital. The Y variable is the number patients discharged during a particular time period. The GREG estimator for this population is based on the model $E_M(Y) = \beta_1 x^{1/2} + \beta_2 x$, $\text{var}_M(Y) = \sigma^2 x$. Samples of size 50 and 100 were selected using simple random sampling without replacement (srswor) and probability proportional to size (pps) without replacement with the size being the square root of x . For each combination of selection method and sample size, 3,000 samples were selected. The estimators \hat{Y}_G , v_π , v_{R1} , v_{R2} , v_{SSW} , v_D , v_{DP} , v_J^* , v_{JP}^* , and v_J were calculated for each sample. For comparison we also included the π -estimator, $\hat{Y}_\pi = \hat{T}_\pi/N$. The variance estimator v_T was included but is not reported here since results were little different from v_{R2} .

The Labor Force population contains 10,841 persons. The auxiliary variables used were age, sex, and number of hours worked per week. The Y variable was total weekly wages. Age was grouped into four categories: 19 years and under, 20-24, 25-34, and 35 or more. The model for the GREG included an intercept, main effects for age and sex, and the quantitative variable, hours worked. A constant model-variance was used. Samples of size 50, 100, and 250 were selected. The two selection methods used were srswor and sampling without replacement with probability proportional to hours worked. (This population has some clustering but this was ignored in these simulations.)

The third population was a version of Labor Force designed to inject some outliers or skewness into the weekly wages variable. We denote this new version as

"LF(mod)" for reference. In the original Labor Force population, weekly wages were top-coded at \$999. For each such top-coded wage, a new wage was generated equal to \$1,000 plus a lognormal random variable whose distribution had scale and shape parameters of 6.9 and 1. Recoded wages were generated for 4.4% of the population. Prior to recoding, the annualized mean wage was \$19,359, and the maximum was \$51,948; after recoding, the mean was \$23,103 and the maximum was \$608,116. Thus, LF(mod) exhibits more of the skewness in income that would be found in a real population.

The resulting LF(mod) distribution is shown in Figure 1 where weekly wages is plotted against hours worked for subgroups defined by age. In each panel the black points are for males while the open circles are for females. A horizontal reference line is drawn in each panel at \$999. Although there is a considerable amount of over-plotting, the general features are clear. Wage levels and spread go up

as age increases, hours worked per week is related, though somewhat weakly, to wages, and wages are most skewed for age groups 25-34 and 35+. Less evident is the fact that wages for males are generally higher than ones for females.

Table 1 shows the empirical percentage relative biases, defined as the average over the samples of $(\hat{T} - T)/T$ for the π -estimator and general regression estimator for the various populations and sample sizes. Root mean square errors (rmse's), defined as the square root of the average over the samples of $(\hat{T} - T)^2$, are also shown. In the Hospitals population, both estimators have negligible bias at either sample size. The GREG is considerably more efficient in Hospitals than the π -estimator because of a strong relationship of Y to x . In the two Labor Force populations, both the π -estimator and the GREG are nearly unbiased while the GREG is somewhat more efficient as measured by the rmse for all sample sizes and selection methods.

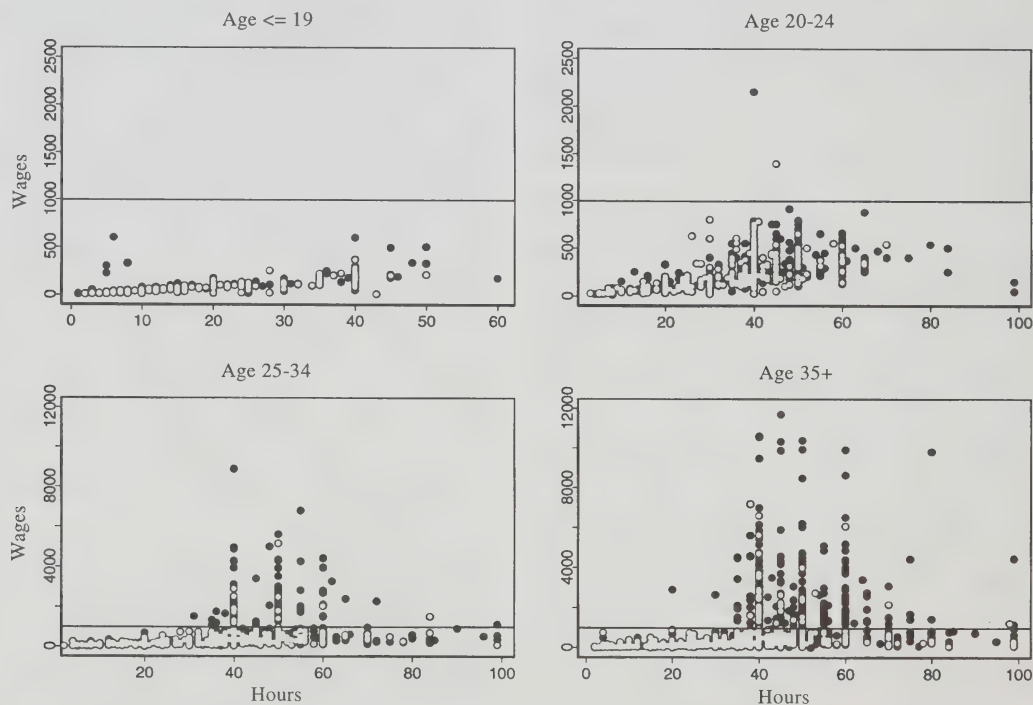


Figure 1. Scatterplots of Weekly Wages versus Hours Worked per Week in Four Age Groups for the LF(mod) population. Open circles are for females. Black circles are for males. A horizontal line is drawn at \$999 per week, the maximum value in the original Labor Force population.

Table 1

Relative biases and root means square errors (rmse's) of the π -estimator and the general regression in different simulation studies of 3,000 samples each.

	Hospitals			Labor Force			LF(mod)		
	<i>n</i> =50	<i>n</i> =100	<i>n</i> =50	<i>n</i> =100	<i>n</i> =250	<i>n</i> =50	<i>n</i> =100	<i>n</i> =250	
Simple random samples									
\hat{Y}_{π}									
Relbias (%)	0.2	-0.1	-0.6	0	0	-0.1	0	-0.3	
rmse	76.6	50.7	34.2	24.1	15.5	88.6	61.2	38.8	
\hat{Y}_G									
Relbias (%)	0.2	0.2	0.1	0.1	0.2	0.4	0.2	-0.1	
rmse	32.6	21.1	28.3	19.9	12.4	86.0	57.4	36.0	
Probability proportional to size samples									
\hat{Y}_{π}									
Relbias (%)	-0.1	0.1	-0.5	0	0	0	-0.1	-0.1	
rmse	37.6	24.4	28.2	20.3	12.6	80.6	54.6	34.1	
\hat{Y}_G									
Relbias (%)	0.1	0.1	-0.10	0.10	0	-0.6	-0.7	-0.4	
rmse	27.2	16.9	28.2	19.3	12.0	81.8	55.1	33.5	

Table 2 lists the empirical relative biases (relbiases) of the nine variance estimators, defined as $100(\bar{v} - \text{mse})/\text{mse}$, where \bar{v} is the average of a variance estimator over the 3,000 samples and mse is the empirical mean square error of the GREG. The rows of the table are sorted by the size of the relbias in LF(mod) for srswor's of size 50, although the ordering would be similar for the other populations, sample sizes, and selection methods. In the Hospitals population, the sampling fraction is substantial, especially when $n=100$. As might be expected, this results in the estimators that omit any type of finite population correction (\hat{fpc})— v_{R2} , v_D , v_J^* , and v_J —being severe over-estimates in either srswor or pps samples. Because v_{R1} lacks a term to reflect the model-variance of the nonsample sum, it underestimates the mse badly when the sampling fraction is large.

In the Labor Force and LF(mod) populations, increasing sample size leads to decreasing bias. The estimators v_π , v_{R1} , v_{R2} , and v_{SSW} have negative biases that tend to be less severe as the sample size increases. The jackknife v_J and its variants, v_J^* , v_{JP} , are over-estimates, especially at $n=50$. The estimators, v_D and v_{DP} , are more nearly unbiased at each of the sample sizes than most of the other estimators.

The empirical coverages of 95% confidence intervals across the 3,000 samples in each set are shown in Table 3 for the Hospitals population. The three choices of variance estimator that use the leverage adjustments but not \hat{fpc} 's— v_D , v_J^* , and v_J —are larger and, thus, have higher coverage rates than v_π , v_{R2} , and v_{SSW} . The tendency of the jackknife to be larger than other variance estimates for the GREG has also been noted by Stukel, Hidiroglou, and Särndal (1996). This is an advantage for the smaller sample size, $n=50$. When $n=100$ and the sampling fraction is large, the estimators with the \hat{fpc} 's— v_π , v_{SSW} , v_{DP} , and v_J^* —have closer to the nominal 95% coverage rates while v_{R2} , v_D , v_J^* , and v_J cover in about 97 or 98% of the samples. The estimator v_{JP} , that approximates the

jackknife but includes an \hat{fpc} , is a good choice at either sample size or sampling plan.

Table 2

Relative biases of nine variance estimators for the general regression estimator in different simulation studies of 3,000 samples each.

	Hospitals			Labor Force			LF(mod)	
	<i>n</i> =50	<i>n</i> =100	<i>n</i> =50	<i>n</i> =100	<i>n</i> =250	<i>n</i> =50	<i>n</i> =100	<i>n</i> =250
Simple random samples								
v_π	-8.6	-4.2	-18.1	-12.3	-7.5	-16.3	-2.8	-2.6
v_{R1}	-18.9	-27.0	-11.3	-9.9	-8.0	-9.6	-0.7	-3.3
v_{SSW}	-7.6	-3.0	-10.9	-9.1	-5.9	-9.3	0.1	-1.1
v_{R2}	5.9	30.1	-10.5	-8.2	-3.7	-8.8	1.0	1.3
v_{DP}	-1.4	0.2	0.1	-3.8	-3.8	0.6	5.1	0.8
v_D	13.0	34.3	0.6	-2.9	-1.6	1.0	6.1	3.2
v_J	18.4	37.4	13.9	2.2	0.3	11.2	10.5	4.8
v_{JP}^*	5.4	3.5	14.0	2.1	-1.7	12.4	10.5	2.7
v_J^*	20.8	38.8	14.5	3.1	0.7	12.9	11.5	5.2
Probability proportional to size samples								
v_π	-5.9	-0.9	-22.1	-12.1	-6.8	-16.5	-10.6	-0.3
v_{R1}	-19.7	-32.4	-11.9	-7.7	-7.1	-9.1	-8.2	-2.7
v_{SSW}	-4.0	0.0	-11.6	-7.0	-4.9	-8.7	-7.3	-0.1
v_{R2}	16.0	52.6	-11.2	-6.0	-2.5	-8.3	-6.3	2.6
v_{DP}	0.1	2.0	0.8	-0.3	-1.6	0.9	-2.5	2.1
v_D	20.8	55.6	1.3	0.7	0.8	1.4	-1.5	4.8
v_J	23.6	57.2	22.6	11.8	5.3	14.6	4.7	7.3
v_{JP}^*	4.4	4.0	19.7	9.3	3.1	14.8	3.9	4.9
v_J^*	26.1	58.8	20.3	10.3	5.5	15.4	5.0	7.7

Table 3

95% confidence interval coverage rates for simulations using the Hospitals population and nine variance estimators. 3,000 simple random samples and probability proportional to size were selected without replacement for samples of size 50 and 100. *L* is percent of samples with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} < -1.96$; *M* is percent with $|\hat{Y}_G - \bar{Y}|/\sqrt{v}^{1/2} \leq 1.96$; *U* is percent with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} < 1.96$.

	<i>n</i> =50			<i>n</i> =100		
	<i>L</i>	<i>M</i>	<i>U</i>	<i>L</i>	<i>M</i>	<i>U</i>
Simple random samples						
v_π	3.1	92.1	4.8	2.6	93.6	3.9
v_{R1}	4.2	91.0	4.7	4.8	89.8	5.5
v_{SSW}	3.3	92.5	4.2	2.8	94.0	3.1
v_{R2}	2.8	93.9	3.3	1.4	97.0	1.6
v_{DP}	3.1	93.0	3.9	2.7	94.3	2.9
v_D	2.4	94.6	3.0	1.2	97.3	1.5
v_J	2.2	95.0	2.8	1.2	97.3	1.5
v_{JP}^*	2.9	93.6	3.5	2.6	94.6	2.9
v_J^*	2.2	95.1	2.8	1.2	97.4	1.4
Probability proportional to size samples						
v_π	2.9	93.9	3.2	2.6	94.6	2.8
v_{R1}	4.1	92.0	3.9	5.0	89.3	5.7
v_{SSW}	2.9	94.2	2.9	2.6	94.8	2.6
v_{R2}	2.1	95.8	2.1	0.9	98.3	0.8
v_{DP}	2.7	94.5	2.8	2.5	95.0	2.5
v_D	1.9	96.2	1.9	0.9	98.3	0.8
v_J	1.8	96.3	1.9	0.9	98.4	0.7
v_{JP}^*	2.6	94.8	2.6	2.4	95.4	2.2
v_J^*	1.7	96.5	1.8	0.8	98.4	0.7

Tables 4 and 5 show the coverage rates for the Labor Force and LF(mod) populations. For the former, v_{DP} , v_D , v_J , v_{JP} , and v_J^* are clearly better in Labor Force at $n=50$ for both srswor and pps samples. But, for $n=250$, coverage rates are similar for all estimators. The purely design-based estimator, v_π , is unsatisfactory at the smaller sample sizes for either sampling plan. As in Hospitals, v_{JP} gives near nominal coverage at each sample size in the Labor Force population.

The most striking results in Tables 4 and 5 are for LF(mod) where all variance estimators give poor coverage. Coverage rates range from 78.0% for the combination (v_π , $n=50$, srswor) to 90.7% for (v_J and v_J^* , $n=250$, pps). Virtually all cases of non-coverage are because $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} < -1.96$, where v is any of the variance estimators. The poor coverage rates occur even though the π -estimator and GREG are unbiased over all samples (see Table 1) and, in the cases of v_J , v_{JP} , and v_J^* , the variance estimators are overestimates (see Table 2).

Table 4

95% confidence interval coverage rates for simulations using the Labor Force and LF(mod) populations and nine variance estimators. 3,000 simple random samples were selected without replacement for samples of size 50 and 100. L is percent of samples with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} < -1.96$; M is percent with $|\hat{Y}_G - \bar{Y}|/\sqrt{v}^{1/2} \leq 1.96$; U is percent with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} > 1.96$.

	n=50			n=100			n=250		
	L	M	U	L	M	U	L	M	U
Labour Force									
v_π	5.3	91.4	3.2	4.3	92.8	2.9	2.8	94.1	3.1
v_{R1}	4.9	92.4	2.7	4.3	93.0	2.7	2.8	93.9	3.3
v_{SSW}	4.9	92.5	2.6	4.3	93.1	2.7	2.8	94.1	3.1
v_{R2}	4.9	92.5	2.6	4.2	93.2	2.6	2.5	94.6	2.9
v_{DP}	4.2	93.6	2.2	3.9	93.7	2.4	2.6	94.5	2.9
v_D	4.2	93.6	2.2	3.9	93.9	2.2	2.4	94.9	2.7
v_J	3.0	95.1	1.9	3.4	94.7	1.9	2.4	95.0	2.7
v_{JP}	3.0	95.1	1.9	3.3	94.7	1.9	2.5	94.8	2.7
v_J^*	3.0	95.1	1.9	3.3	94.8	1.9	2.4	95.0	2.7
LF(mod)									
v_π	21.0	78.0	0.9	14.1	85.5	0.4	9.9	89.7	0.4
v_{R1}	20.9	78.7	0.3	14.1	85.7	0.2	10.2	89.5	0.3
v_{SSW}	20.9	78.8	0.3	14.0	85.8	0.2	9.9	89.9	0.3
v_{R2}	20.8	78.8	0.3	13.8	86.0	0.2	9.7	90.1	0.3
v_{DP}	19.7	80.0	0.2	13.4	86.5	0.1	9.7	90.1	0.3
v_D	19.7	80.0	0.2	13.2	86.7	0.1	9.6	90.1	0.3
v_J	18.4	81.4	0.2	12.7	87.2	0.1	9.4	90.3	0.3
v_{JP}	18.4	81.4	0.2	12.7	87.2	0.1	9.5	90.2	0.3
v_J^*	18.3	81.5	0.2	12.6	87.3	0.1	9.3	90.4	0.3

Negative estimation errors, $\hat{Y}_G - \bar{Y}$, occur in samples that include relatively few persons with large weekly wages. Figure 2 is a plot of t -statistics based on $\sqrt{v_{JP}}$, i.e., $(\hat{Y}_G - \bar{Y})/\sqrt{v_{JP}}$, versus the number of sample persons with weekly wages of \$1,000 or more in sets of 1,000 samples for (srswor; $n=50, 100, 250$). The negative estimation errors in samples with few persons with high incomes lead to negative t -statistics, and confidence intervals that miss the population mean on the low side. The problem decreases with increasing sample size, but the convergence

to the nominal coverage rates is slow and occurs "from the bottom up." Regardless of the variance estimator used, coverage will be less than 95% unless the sample is quite large.

Table 5

95% confidence interval coverage rates for simulations using the Labor Force and LF(mod) populations and nine variance estimators. 3,000 probability proportional to size samples were selected without replacement for samples of size 50, 100 and 250. L is percent of samples with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} < -1.96$; M is percent with $|\hat{Y}_G - \bar{Y}|/\sqrt{v}^{1/2} \leq 1.96$; U is percent with $(\hat{Y}_G - \bar{Y})/\sqrt{v}^{1/2} > 1.96$.

	n=50			n=100			n=250		
	L	M	U	L	M	U	L	M	U
Labour Force									
v_π	5.7	90.2	4.1	3.7	92.9	3.4	3.1	94.3	2.6
v_{R1}	5.3	92.1	2.6	3.3	93.8	2.9	3.5	94.0	2.5
v_{SSW}	5.2	92.2	2.6	3.2	94.0	2.9	3.3	94.4	2.2
v_{R2}	5.2	92.3	2.6	3.1	94.1	2.8	3.0	94.8	2.2
v_{DP}	4.3	93.6	2.0	2.9	94.7	2.4	3.0	94.9	2.1
v_D	4.3	93.7	2.0	2.9	94.7	2.4	2.8	95.1	2.1
v_J	3.3	95.5	1.2	2.4	95.8	1.7	2.6	95.5	1.9
v_{JP}	3.3	95.4	1.3	2.6	95.5	1.9	2.7	95.3	1.9
v_J^*	3.3	95.4	1.3	2.6	95.6	1.8	2.6	95.6	1.8
LF(mod)									
v_π	19.6	79.7	0.7	15.0	84.4	0.7	9.9	89.8	0.4
v_{R1}	20.2	79.6	0.2	15.9	83.8	0.3	10.3	89.4	0.3
v_{SSW}	20.1	79.7	0.2	15.8	84.0	0.3	10.0	89.8	0.2
v_{R2}	20.1	79.7	0.2	15.6	84.1	0.2	9.8	90.0	0.2
v_{DP}	18.7	81.1	0.2	14.8	85.0	0.1	9.7	90.0	0.2
v_D	18.7	81.1	0.2	14.7	85.2	0.1	9.4	90.4	0.2
v_J	16.6	83.2	0.1	13.6	86.4	0.0	9.1	90.7	0.2
v_{JP}	16.6	83.3	0.1	13.9	86.1	0.0	9.4	90.4	0.2
v_J^*	16.5	83.4	0.1	13.8	86.2	0.0	9.1	90.7	0.2

We also examined how well the variance estimators perform, conditional on sample characteristics. We present only results related to bias of the variance estimators to conserve space. For the Hospitals population, we sorted the samples based on $D_x = \mathbf{1}'(\hat{\mathbf{T}}_x - \mathbf{T}_x)$, which is the sum of the differences of the π -estimates of the totals of $x^{1/2}$ and x from their population totals. Twenty groups of 150 samples each were then formed. In each group, we computed the bias of \hat{Y}_G along with the rmse, and the square root of the average of each variance estimator. The results are plotted in Figure 3 for srswor with $n=50$ and 100 and for pps with $n=50$ and 100. A subset of the variance estimators is plotted. The horizontal axis in each panel gives values of D_x . Since v_J , v_J^* , v_D , and v_{R2} are similar through most of the range of D_x , only the jackknife v_J is plotted. Also, v_{DP} and v_{JP} are close, and only the latter is plotted. The GREG does have a conditional bias that affects the rmse in off-balance samples. The poor conditional properties of v_π are most evident in the simple random samples where the bias of v_π as an estimate of the mse runs from negative to positive over the range of D_x . Among the other variance estimates, conditional biases are similar to the unconditional biases in Table 2. Both v_{JP} and v_{SSW} are in theory approximately design and model-unbiased, and both track the rmse well.

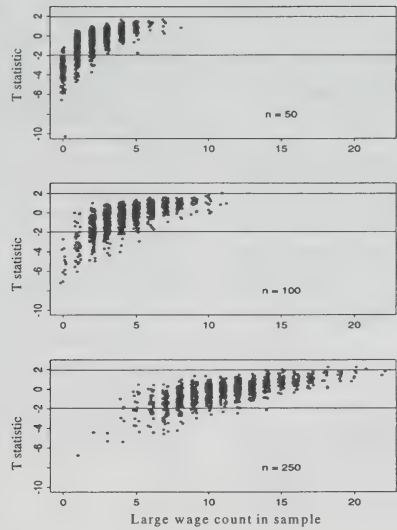


Figure 2. Plot of t -statistics versus the number of sample persons with weekly wages greater than \$1,000 in the sets of 1,000 simple random samples of size $n = 50, 100, 250$ from the LF(mod) population. Horizontal reference lines are drawn at ± 1.96 . Points are jittered to minimize overplotting.

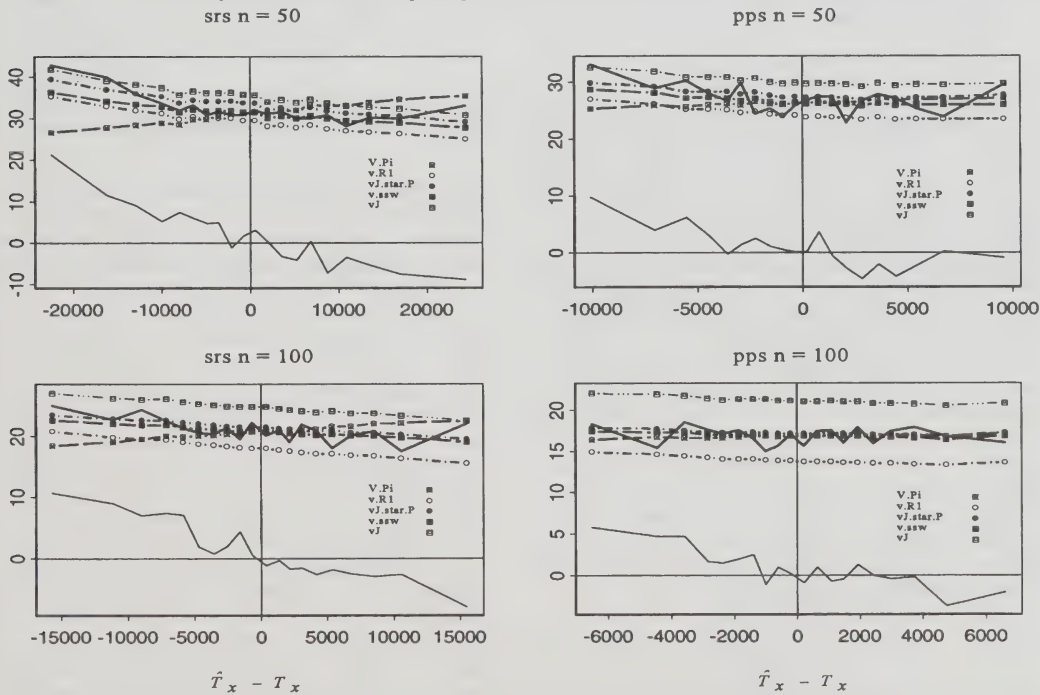


Figure 3. Plot of conditional biases, rmse's, and means of standard error estimates of the GREG for the samples from the Hospitals population. Horizontal and vertical reference lines are drawn at 0. The lowest curve each panel is the bias of the GREG. The thick solid line is the conditional root mean square error.

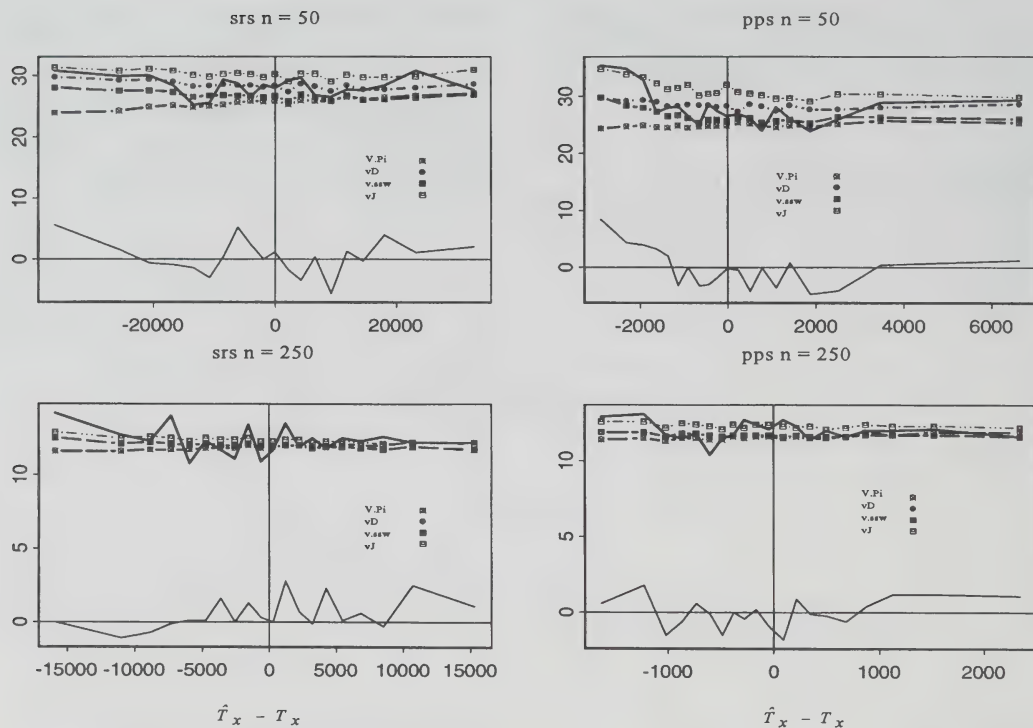


Figure 4. Plot of conditional biases, rmse's, and means of standard error estimates of the GREG for the samples from the Labor Force population. Horizontal and vertical reference lines are drawn at 0. The lowest curve in each panel is the bias of the GREG. The thick solid line is the conditional root mean square error.

Figure 4 is a similar plot for the samples from the Labor Force population. The following sets of estimates are very similar and only the first in each set is included in the plots: (v_D, v_{DP}) , and (v_J, v_J^*, v_{JP}) . Only the srswor and pps samples of size $n = 50$ and 250 are included. The horizontal axis is again D_x , which is the sum of differences between the π -estimates and the population values of the totals for age and sex groups and the number of hours worked per week. The conditional bias of v_π is evident in samples with the smallest values of D_x but the problem diminishes for the larger sample size in both srswor and pps samples. The jackknife v_J is, on average, the largest of the variance estimators throughout the range of D_x . The differences among the variance estimates and their biases are less for the larger sample size. The estimators v_D , v_{SSW} , and v_J all track the rmse reasonably well except when D_x is most negative, where all are somewhat low.

5. CONCLUSION

A variety of estimators of the variance of the general regression estimator have been proposed in the sampling literature, mainly with the goal of estimating the design-based variance. Estimators can be easily constructed that

are approximately unbiased for both the design-variance and, under certain models, the model-variance. Moreover, the dual-purpose estimators studied here are robust estimators of a model-variance even if the model that motivates the GREG has an incorrect variance parameter.

A key feature of the best of these estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis. The desirability of using leverage corrections to regression variance estimators in order to combat heteroscedasticity is well-known in econometrics, having been proposed by MacKinnon and White (1985) and recently revisited by Long and Ervin (2000). One of the best choices is an approximation to the jackknife, denoted here by v_{JP}^* , that includes a type of finite population correction.

The robust estimators studied here are quite useful for variables whose distributions are reasonably "well behaved." They adjust variance estimators in small and moderate size samples in a way that often results in better confidence interval coverage. However, they are no defense when variables are extremely skewed, and large observations are not well represented in a sample. Whether one refers to this problem as one of skewness or of outliers, the effect is clear. A sample that does not include a sufficient

number of units with large values will produce an estimated mean that is too small. A variance estimator that is small often accompanies the small estimated mean. As the simulations in section 4 illustrate, in such samples even the best of the proposed variance estimators will not yield confidence intervals that cover at the nominal rate. The transformation methods of Chen and Chen (1996) might hold some promise, but that approach would have to be tested for the more complex GREG estimators studied here.

The most effective solution to the skewness problem does not appear to be to make better use of the sample data. Rather, the sample itself needs to be designed to include good representation of the large units. In many cases, however, like a survey of households to measure income or capital assets, this may be difficult or impossible if auxiliary information closely related to the target variable is not available. Better use of the sample data employing models for skewed variables may then be useful (see, e.g., Karlberg 2000).

ACKNOWLEDGEMENT

The author is indebted to Alan Dorfman whose ideas were the impetus for this work and to the Associate Editor and two referees for their careful reviews.

APPENDIX: Details of Jackknife Calculations

Using (3.2), (3.3), and the standard matrix result in Lemma 5.4.1 of Valliant *et al.* (2000), we have

$$\mathbf{A}_{\pi s(i)}^{-1} = \left[\mathbf{A}_{\pi s}^{-1} + \frac{\mathbf{A}_{\pi s}^{-1} \mathbf{x}_i \mathbf{x}_i' \mathbf{A}_{\pi s}^{-1} / v_i \pi_i}{1 - h_{ii}} \right].$$

From this and the definition of $\hat{\mathbf{B}}_{(i)}$, the slope estimator, omitting unit i , is $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} + n^{-1} \sum_s \mathbf{Q}_i$ where

$$\mathbf{Q}_i = \frac{\mathbf{A}_{\pi s}^{-1} \mathbf{x}_i}{1 - h_{ii}} \frac{r_i}{v_i \pi_i}.$$

The GREG estimator, after deleting unit i , is

$$\hat{T}_{G(i)} = \frac{n}{n-1} \left(\hat{T}_\pi - \frac{Y_i}{\pi_i} \right) + (\hat{\mathbf{B}} - \mathbf{Q}_i)' \left[\mathbf{T}_x - \frac{n}{n-1} \left(\hat{\mathbf{T}}_x - \frac{\mathbf{x}_i}{\pi_i} \right) \right].$$

After some rearrangement, this can be rewritten as

$$\hat{T}_{G(i)} = \frac{n}{n-1} \hat{T}_G - \frac{n}{n-1} \left[\frac{g_i r_i}{\pi_i (1 - h_{ii})} \right] + \frac{n}{n-1} G_i + \frac{1}{n-1} K_i$$

where

$$G_i = \frac{h_{ii} Y_i - \hat{Y}_i}{\pi_i (1 - h_{ii})}$$

and

$$K_i = (\hat{\mathbf{B}} - \mathbf{Q}_i)' \left(\frac{n \mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x \right).$$

It follows that $\hat{T}_{G(i)} - \hat{T}_{G(i)} = -n(n-1)^{-1} (D_i - \bar{D}_s) + n(n-1)^{-1} F_i$ where $F_i = (G_i - \bar{G}_s) + n^{-1} (K_i - \bar{K}_s)$ with \bar{G}_s and \bar{K}_s being sample means with the obvious definitions. Substituting in the jackknife formula (3.1) gives

$$v_J(\hat{Y}_G) = N^{-2} \frac{n}{n-1} \times \left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2 \sum_s F_i (D_i - \bar{D}_s) \right]. \quad (\text{A.1})$$

Formula (A.1) is exact, but with some further approximations we can get the relative sizes of the terms. Using the values of G_i and K_i above and the fact that h_{ii} and the elements of \mathbf{Q}_i are $o(1)$, we have

$$\begin{aligned} G_i + n^{-1} K_i &= \frac{h_{ii} Y_i - \hat{Y}_i}{\pi_i (1 - h_{ii})} + \frac{1}{n} (\hat{\mathbf{B}} - \mathbf{Q}_i)' \left(\frac{n \mathbf{x}_i}{\pi_i} - \hat{\mathbf{T}}_x \right) \\ &\approx -\frac{\hat{Y}_i}{\pi_i} + \hat{\mathbf{B}}' \frac{\mathbf{x}_i}{\pi_i} - \frac{1}{n} \hat{\mathbf{B}}' \hat{\mathbf{T}}_x \\ &= -\frac{1}{n} \hat{\mathbf{B}}' \hat{\mathbf{T}}_x \end{aligned}$$

where \approx denotes "asymptotically equivalent to." It follows that $F_i \approx 0$ and that $v_J(\hat{Y}_G) \approx \sum_s (D_i - \bar{D}_s)^2$, i.e., (3.5) holds.

Next, we can show that the second term in (3.5) converges in probability to zero. The vector of residuals can be expressed as $\mathbf{r}_s = (\mathbf{I} - \mathbf{H}_s) \mathbf{Y}_s$, and the second term in (3.5) is equal to $N^{-2} n^{-1} \mathbf{g}_s' \Pi_s^{-1} \mathbf{U}^{-1} \mathbf{r}_s \mathbf{r}_s' \mathbf{U}^{-1} \Pi_s^{-1} \mathbf{g}_s$ where $\mathbf{U} = \text{diag}(1 - h_{ii})$, $i \in s$. Thus, the second term in (3.5) is the square of $\mathbf{B} = N^{-1} n^{-1/2} \mathbf{g}_s' \Pi_s^{-1} \mathbf{U}^{-1} \mathbf{r}_s$ which has expectation zero under any model with $E_M(\mathbf{r}_i) = 0$. The model-variance of \mathbf{B} is

$$\begin{aligned} N^{-2} n^{-1} \text{var}_M(\mathbf{g}_s' \Pi_s^{-1} \mathbf{U}^{-1} \mathbf{r}_s) &= \\ N^{-2} n^{-1} \mathbf{g}_s' \Pi_s^{-1} \mathbf{U}^{-1} (\mathbf{I} - \mathbf{H}_s) \times & \quad (\text{A.2}) \\ \mathbf{V}_s (\mathbf{I} - \mathbf{H}_s)' \mathbf{U}^{-1} \Pi_s^{-1} \mathbf{g}_s & \end{aligned}$$

which has order of magnitude n^{-2} under the assumptions we have made. Consequently, the second term in (3.5) is the square of a term with mean zero and a model-variance that approaches zero as the sample size increases. The second term in (3.5) then converges to zero by Chebyshev's inequality. This justifies (3.6).

REFERENCES

- BELSLEY, D.A., KUH, E. and WELSCH, R.E. (1980). *Regression Diagnostics*. New York: John Wiley & Sons, Inc.
- BREWER, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 in *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College, and P.S. Kott). New York: John Wiley & Sons, Inc., 589-606.
- BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 205-212.
- CHEN, G., and CHEN, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.
- DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.
- HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1980). SUPERCARP. Department of Statistics. Ames, Iowa: Iowa State University.
- KARLBERG, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-243.
- KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.
- KREWSKI and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LONG, J.S., and ERVIN, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54, 217-224.
- MACKINNON, J.G., and WHITE, H. (1985). Some heteroskedastic consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 53-57.
- RAO, J.N.K., and WU, C.J.F. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SÄRNDAL, C.-E., and WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- STUKEL, D., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.
- SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- VALLIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.
- YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.
- YUNG, W., and RAO, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 18, No. 1, 2002

Evaluating Socio-economic (SES) Bias in Survey Nonresponse John Goyder, Keith Warriner, and Susan Miller	1
Are Nonrespondents to Health Surveys Less Healthy than Respondents G. Cohen and J.C. Duffy	13
Accounting for Biases in Election Surveys; The Case of the 1998 Quebec Election Claire Durand, Andre Blais, and Sebastien Vachon	25
Small Area Estimation via Generalized Linear Models Alastair Noble, Stephen Haslett, and Greg Arnold	45
Generalized Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI) Jan de Haan	61
Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation Risto Lehtonen, Erkki Pahkinen, and Carl-Erik Särndal	87
Book and Software Reviews	111

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

Richard A. LOCKHART Editor's Report/Rapport du rédacteur en chef	1
Yong YOU and J.N.K. RAO Small area estimation using unmatched sampling and linking models	3
Debbie J. DUPUIS and Stephan MORGENTHALER Robust weighted likelihood estimators with an application to bivariate extreme value problems	17
Patrick BÉLISLE, Lawrence JOSEPH, David B. WOLFSON and Xiaojie ZHOU Bayesian estimation of cognitive decline in patients with Alzheimer's disease	37
Joseph G. IBRAHIM, Ming-Hui CHEN and Stuart R. LIPSITZ Bayesian methods for generalized linear models with covariates missing at random	55
Mario TROTTINI and Fulvio SPEZZAFERRI A generalized predictive criterion for model selection	79
Pamela OHMAN-STRICKLAND and George CASELLA Approximate and estimated saddlepoint approximations	97
Yong B. LIM, Jerome SACKS, W.J. STUDDEN and William J. WELCH Design and analysis of computer experiments when the output is highly correlated over the input space	109
Boxin TANG, Fengshi MA, Debra INGRAM and Hong WANG Bounds on the maximum number of clear two-factor interactions for 2^{m-p} designs of resolution III and IV	127
<i>Case study in data analysis: The genetic analysis of inflammatory bowel disease</i>	137
Lucia MIREA, Shelley B. BULL, Mark S. SILVERBERG and Katherine A. SIMINOVITCH <i>Introduction and Analysis I: The genetic analysis of a complex disease</i>	138
Jiahua CHEN, John D. KALBFLEISCH and Sandra ROMERO-HIDALGO <i>Analysis 2: Genetic data analysis of affected sib pairs</i>	145
Gerarda A. DARLINGTON and Andrew D. PATERSON <i>Analysis 3: Genetic analysis of chromosome 6 in inflammatory bowel disease</i>	152
Nicole M. ROSLIN, J.C. LOREDO-OSTI, Celia M.T. GREENWOOD and Kenneth MORGAN <i>Analysis 4: Genetic analysis of the role of the HLA region in inflammatory bowel disease</i>	158
Christopher A. FIELD and Bruce SMITH Discussion of the evaluation of a candidate genetic locus in a genome scan of complex disease	167
Acknowledgement of referees' service/Remerciements aux membres des jurys	175
Forthcoming Papers/Articles à paraître	176

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préféablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; l, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 Exemple: Cochran (1977, p. 164).
La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

1	Richard A. LOCKHART Editor's Report/Rapport du rédacteur en chef
3	Yong YOU and J.N.K. RAO Small area estimation using unmatched sampling and linking models
17	Debbie J. DUPUIS and Stephan MORGENTHAUER Robust weighted likelihood estimators with an application to bivariate extreme value problems
37	Patrick BELISLE, Lawrence JOSEPH, David B. WOLFSON and Xiaojie ZHOU Bayesian estimation of cognitive decline in patients with Alzheimer's disease
55	Joseph G. IBRAHIM, Ming-Hui CHEN and Stuart R. LIPSITZ Bayesian methods for generalized linear models with covariates missing at random
79	Mario TROTTINI and Fulvio SPEZZAFERRI A generalized predictive criterion for model selection
97	Pamela OHMAN-STICKLAND and George CASELLA Approximate and estimated saddlepoint approximations
109	Yong B. LIM, Jerome SACKS, W.J. STUDDEN and William J. WELCH Design and analysis of computer experiments when the output is highly correlated over the input space
127	Boxin TANG, Fengshi MA, Debra INGRAM and Hong WANG Bounds on the maximum number of clear two-factor interactions for 2^{m-p} designs of resolution III and IV
137	Case study in data analysis: The genetic analysis of inflammatory bowel disease
138	Lucia MIREA, Shelley B. BULL, Mark S. SILVERBERG and Katherine A. SIMINOVITCH Introduction and Analysis 1: The genetic analysis of a complex disease
145	Jiahua CHEN, John D. KALBFLEISCH and Sandra ROMERO-HIDALGO Analysis 2: Genetic data analysis of affected sib pairs
152	Gerarda A. DARLINGTON and Andrew D. PATERSON Analysis 3: Genetic analysis of chromosome 6 in inflammatory bowel disease
158	Nicole M. ROSLIN, J.C. LOREDO-OSTI, Celia M.T. GREENWOOD and Kenneth MORGAN Analysis 4: Genetic analysis of the role of the HLA region in inflammatory bowel disease
167	Christopher A. FIELD and Bruce SMITH Discussion of the evaluation of a candidate genetic locus in a genome scan of complex disease
175	Acknowledgement of referees' service/Remerciements aux membres des jurys
176	Forthcoming Papers/Articles à paraître

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 18, No. 1, 2002

Evaluating Socio-economic (SES) Bias in Survey Nonresponse John Goyder, Keith Warriner, and Susan Miller	1
Are Nonrespondents to Health Surveys Less Healthy than Respondents G. Cohen and J.C. Duffy	13
Accounting for Biases in Election Surveys: The Case of the 1998 Quebec Election Claire Durand, Andre Blais, and Sebastien Vachon	25
Small Area Estimation via Generalized Linear Models Alastair Noble, Stephen Haslett, and Greg Arnold	45
Generalized Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index (CPI) Jan de Haan	61
Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation Risto Lehtonen, Erkki Pankkinen, and Carl-Erik Särndal	87
Book and Software Reviews	111

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

BIBLIOGRAPHIE

- Valliant : Estimation de la variance de l'estimateur de régression généralisée
- ROYALL, R.M., et CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- ROYALL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SÄRNDAL, C.-E., et WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- STUKEL, D., HIDROGLOU, M.A. et SÄRNDAL, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.
- SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- VALLIANT, R., DORFMAN, A.H. et ROYAL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.
- YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.
- YUNG, W., et RAO, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification. *Journal of the American Statistical Association*, 95, 903-915.
- RAO, J.N.K., et WU, C.J.F. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- MACKINNON, J.G., et WHITE, H. (1985). Some heteroskedastic consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 53-57.
- LONG, J.S., et ERVIN, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54, 217-224.
- REPSKI, R.A., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.
- KARLBERG, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-243.
- SUPERCARP, D., FULLER, W.A. et HICKMAN, R.D. (1980). *Department of Statistics*. Ames, Iowa: Iowa State University.
- DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.
- CHEN, G., et CHEN, J. (1996). Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.
- BREWER, K.R.W. (1999). Le calage esthéticien dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 231-239.
- BREWER, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 dans *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott). New York: John Wiley & Sons, Inc., 589-606.
- BELSLBY, D.A., KUH, E. et WELSCH, R.E. (1980). *Regression Diagnostics*. New York: John Wiley & Sons, Inc.

échantillonales. Il semble plutôt que le plan d'échantillonnage proprement dit doit être conçu de façon à assurer une bonne représentation des grandes unités. Toutefois, dans de nombreux cas, comme celui d'une enquête auprès des ménages afin de mesurer le revenu ou les biens d'équipement, cette solution peut être difficile, voir impossible, si l'on ne dispose pas de données auxiliaires étroitement liées à la variable étudiée. Le cas échéant, il pourrait être utile de mieux utiliser les données échantillonales au moyen de modèles pour variables asymétriques (voir, par exemple Karlborg 2000).

REMERCIEMENTS

L'auteur remercie Alan Dorfman dont les idées sont à l'origine de la présente étude, ainsi que le rédacteur en chef adjoint et les deux évaluateurs pour leur examen minutieux.

ANNEXE : Détails des calculs de l'estimateur Jackknife

L'utilisation de (3.2), (3.3) et du résultat de la matrice type dans le lemme 5.4.1 de Valliant et coll. (2000) nous donne

$$\mathbf{A}_{\pi_s(i)}^{-1} = \left[\mathbf{A}_{\pi_s}^{-1} + \frac{\mathbf{A}_{\pi_s}^{-1} \mathbf{x}_i' \mathbf{x}_i' \mathbf{A}_{\pi_s}^{-1}}{\mathbf{A}_{\pi_s}^{-1} \mathbf{v}_i' \mathbf{v}_i} \right]^{-1}$$

De ceci et de la définition de $\hat{\mathbf{B}}^{(i)}$, il découle que l'estimateur de la pente, si l'on omet l'unité i , est

$$\mathbf{Q}_i = \mathbf{A}_{\pi_s \mathbf{x}_i'} \frac{1 - h_{ii}}{\mathbf{r}_i' \mathbf{v}_i}$$

L'estimateur GREC, après suppression de l'unité i , est

$$\hat{f}_{G(i)} = \frac{n-1}{n} \left(\hat{f}_{\pi} - \frac{\pi_i}{Y_i} \right) + \left(\hat{\mathbf{B}} - \mathbf{Q}_i \right)' \left[\mathbf{T}^x - \frac{n-1}{n} \left(\hat{\mathbf{T}}^x - \frac{\pi_i}{\mathbf{x}_i'} \right) \right]$$

Après certains réarrangements, nous pouvons le réécrire sous la forme

$$\hat{f}_{G(i)} = \frac{n-1}{n} \hat{f}_G - \frac{n-1}{n} \left[\frac{\pi_i (1 - h_{ii})}{\mathbf{r}_i' \mathbf{v}_i} \right] + \frac{n-1}{n} G_i + \frac{1}{n-1} K_i$$

où

$$G_i = \frac{h_{ii} Y_i - \pi_i}{\pi_i (1 - h_{ii})}$$

et

$v_j(\hat{Y}_G) = N^{-2} \frac{n-1}{n} \times$

$$\left[\sum_s (D_i - \bar{D}_s)^2 + \sum_s F_i^2 - 2 \sum_s F_i (D_i - \bar{D}_s) \right] \cdot \quad (\text{A.1})$$

La formule (A.1) est exacte, mais, grâce à certaines approximations supplémentaires, nous pouvons obtenir les tailles relatives des termes. Partant des valeurs de G_i et K_i données plus haut et du fait que h_{ii} et les éléments de \mathbf{Q}_i sont $o(1)$, nous avons

$$G_i + n^{-1} K_i = \frac{\pi_i (1 - h_{ii})}{h_{ii} Y_i - \pi_i} + \frac{1}{n} (\hat{\mathbf{B}} - \mathbf{Q}_i)' \left(\frac{\pi_i}{n \mathbf{x}_i'} - \hat{\mathbf{T}}^x \right) \approx \frac{\pi_i}{Y_i} + \hat{\mathbf{B}}' \frac{\pi_i}{\mathbf{x}_i'} - \frac{1}{n} \hat{\mathbf{B}}' \hat{\mathbf{T}}^x \approx -\frac{1}{n} \hat{\mathbf{B}}' \hat{\mathbf{T}}^x$$

où \approx signifie « asymptotiquement équivalent à ». Il s'ensuit que $F_i \approx 0$ et que $v_j(\hat{Y}_G) \approx \sum_s (D_i - \bar{D}_s)^2$, c'est-à-dire que (3.5) est vérifié. Ensuite, nous pouvons montrer que le deuxième terme de (3.5) converge en probabilité vers zéro. Le vecteur des résidus peut être exprimé sous la forme $\mathbf{r}_s = (\mathbf{I} - \mathbf{H}_s) \mathbf{Y}_s$ et le deuxième terme de (3.5) est égal à $N^{-2} n^{-1} \mathbf{g}_s' \Pi_s^{-1} \mathbf{r}_s' \mathbf{U}_s^{-1} \Pi_s^{-1} \mathbf{g}_s$ où $\mathbf{U} = \text{diag}(1 - h_{ii})$, $\mathbf{B} = N^{-1} n^{-1/2} \mathbf{g}_s' \Pi_s^{-1} \mathbf{U}_s^{-1} \mathbf{r}_s$ dont l'espérance est nulle dans les conditions de tout modèle pour lequel $E_M(\mathbf{r}_i) = 0$. La variance de \mathbf{B} due au modèle est

$$N^{-2} n^{-1} \text{var}_M(\mathbf{g}_s' \Pi_s^{-1} \mathbf{U}_s^{-1} \mathbf{r}_s) =$$

$$N^{-2} n^{-1} \mathbf{g}_s' \Pi_s^{-1} (\mathbf{I} - \mathbf{H}_s) \times$$

$$\mathbf{V}_s (\mathbf{I} - \mathbf{H}_s)' \mathbf{U}_s^{-1} \Pi_s^{-1} \mathbf{g}_s \quad (\text{A.2})$$

dont l'ordre de grandeur est n^{-2} étant donné les hypothèses que nous avons formulées. Par conséquent, le deuxième terme de (3.5) est le carré d'un terme dont la moyenne est nulle et dont la variance due au modèle tend vers zéro à mesure qu'augmente la taille de l'échantillon. Alors, le deuxième terme de (3.5) converge vers zéro conformément à l'inégalité de Chebyshev. Ceci justifie (3.6).

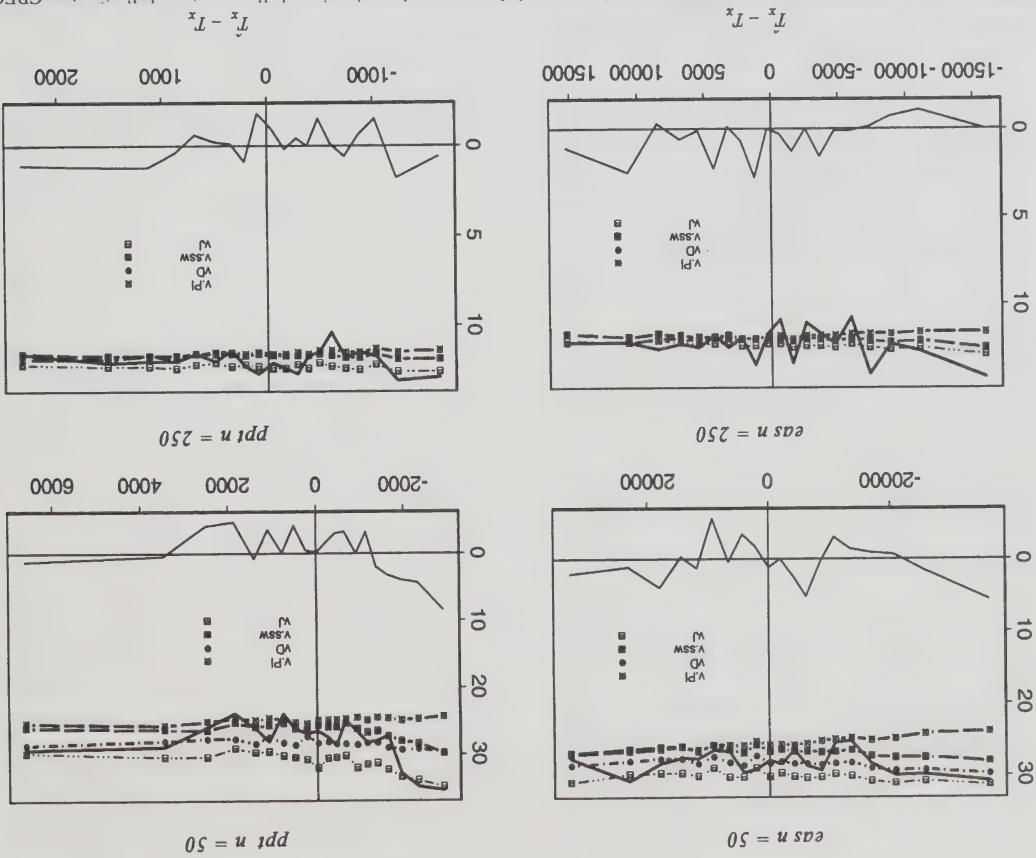


Figure 4. Tracé des valeurs du biais conditionnel, de la régression et de la moyenne des estimations de l'erreur quadratique moyenne conditionnelle. Dans chaque panneau, la courbe la plus basse représente le biais de l'estimateur GREG. La droite en trait plein épais représente la racine de l'erreur quadratique moyenne conditionnelle.

Une caractéristique importante des meilleurs de ces estimateurs est l'ajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Il est bien connu, en économétrie, qu'il est souhaitable de corriger pour les effets leviers les estimateurs par régression de la variance, en vue de combattre l'hétéroscédasticité. Cette correction a été proposée par Mackinnon et White (1985) et réexaminée récemment par Long et Ervin (2000). L'un des meilleurs choix est une approximation de l'estimateur jackknife, représentée ici par v_{jp}^* , qui inclut une forme de correction pour tenir compte de la population finie.

Les estimateurs robustes étudiés ici sont assez utiles pour les variables dont la distribution présente un « comportement » raisonnable. Si les échantillons sont de petite taille ou de taille moyenne, ils produisent des estimateurs corrigés de la variance souvent caractérisés par une meilleure couverture de l'intervalle de confiance. Cependant, ils

n'offrent aucune défense si les variables sont très asymétriques et que les observations dont la valeur est forte ne sont pas bien représentées dans l'échantillon. Qu'il s'agisse d'un problème d'asymétrie ou de valeurs aberrantes, l'effet est évident. Un échantillon ne contenant pas un nombre suffisant d'unités pour lesquelles les valeurs sont élevées produira une estimation trop faible de la variance produisant une estimation trop faible. Comme l'illustreront les simulations décrites à la section 4, pour les échantillons de ce genre, même le meilleur des estimateurs proposés de la variance ne produira pas d'intervalle de confiance dont la couverture correspond au taux nominal. Les méthodes de transformation de Chen et Chen (1996) paraissent prometteuses, mais leur application devrait être testée pour les estimateurs GREG plus complexes étudiés ici.

La résolution la plus efficace du problème d'asymétrie ne consiste pas à mieux utiliser les données

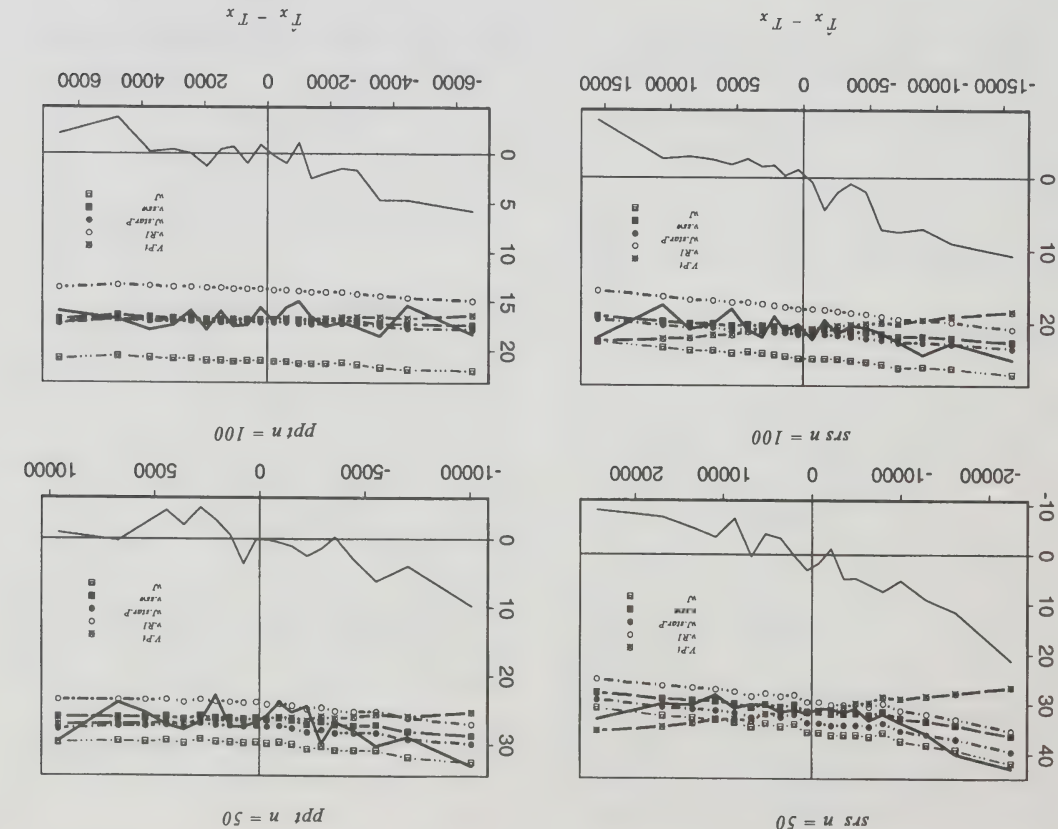


Figure 3. Tracé des valeurs du biais conditionnel, de la regm et de la moyenne des estimations de l'erreur-type de l'estimateur GREG pour les échantillons provenant de la population des hôpitaux. La courbe la plus basse représente le biais de l'estimateur GREG. La droite en trait plein épais représente la racine de l'erreur quadratique moyenne conditionnelle.

La figure 4 présente un graphique similaire pour les échantillons tirés de la population active. Les ensembles d'estimations (v_l , v_k) et (v_l , v_k) sont fort semblables et seule la première de chaque ensemble est incluse dans les graphiques. Seuls les échantillons ppt de taille $n = 50$ et 250 sont inclus. De nouveau, l'axe horizontal est D_x , qui représente la somme des différences entre les estimations π et les valeurs de population des totaux pour les groupes âge-sexe et le nombre d'heures travaillées par semaine. Le biais conditionnel de v_l est évident dans le cas des échantillons pour lesquels les valeurs de D_x sont les plus faibles, mais le problème diminue lorsque la taille d'échantillon augmente, aussi bien dans le cas de l'échantillonnage ppt. L'estimateur jackknife v_j est, en moyenne, celui qui produit l'estimation de la variance la plus importante sur toute l'étendue de D_x . Les différences entre les estimations de la variance et leur biais sont moindres pour les échantillons de

5. CONCLUSION

Divers estimateurs de la variance de l'estimateur de régression généralisée sont proposés dans la littérature sur l'échantillonnage, principalement dans le but d'estimer la variance due au plan de sondage. Il est facile de construire des estimateurs approximativement non biaisés de la variance due au plan de sondage et, dans le cas de certains modèles, de la variance due au modèle. En outre, les estimateurs bivariés étudiés ici sont des estimateurs robustes de la variance due au modèle, même si le modèle qui sert de fondement à l'estimateur GREG contient un paramètre de variance incorrect.

Les estimateurs v_l , v_k et v_lk suivent tous raisonnablement bien la regm, sauf dans les cas où D_x est la plus négative, la valeur étant alors assez faible pour tous les estimateurs.

Des droites horizontale et verticale de référence sont tracées au point zéro. Dans chaque panneau, la courbe la plus basse représente le biais de l'estimateur GREG. La droite en trait plein épais représente la

population du côté de la borne inférieure. La gravité du problème diminue à mesure qu'augmente la taille de l'échantillon, mais la convergence vers le taux de couverture nominal est lente et se produit « de façon ascendante ». Quel que soit l'estimateur de la variance utilisé, la couverture est inférieure à 95 % à moins que l'échantillon ne soit assez grand.

Nous évaluons aussi la performance des estimateurs de la variance en fonction des caractéristiques d'échantillon. Par souci de concision, nous présentons uniquement les résultats concernant le biais des estimateurs de la variance. Dans le cas de la population d'hôpitaux, nous avons trié les échantillons en fonction de $D^x = 1'(\mathbf{T}^x - \mathbf{T})x$, qui est la somme des différences entre les estimateurs π des totaux de $x^{1/2}$ et x et les totaux de population correspondants. Puis, nous avons formé 20 groupes contenant chacun 150 échantillons. Dans chaque groupe, nous avons calculé le biais de \hat{Y}_G , ainsi que la racine de l'erreur quadratique moyenne (regm) et la racine carrée de la moyenne de chaque estimateur de la variance. Les résultats sont présentés graphiquement à la

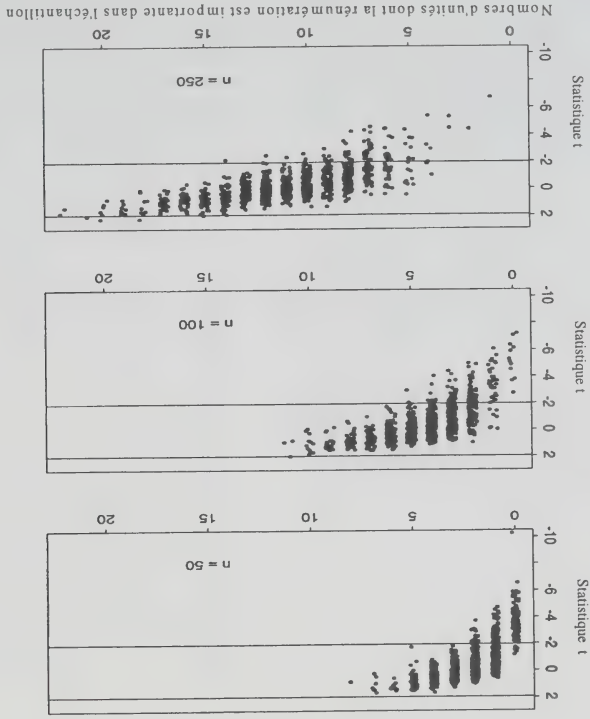


Figure 2. Tracé de la statistique t en fonction du nombre de personnes dans l'échantillon dont la rénumération hebdomadaire est supérieure à 1 000\$ dans les ensembles de 1 000 échantillons aléatoires simples de taille $n = 50, 100, 250$ provenant de la population PA(mod.). Des droites horizontales de référence sont tracées à $\pm 1,96$. Les points sont déplacés légèrement par rapport à leur position normale afin de réduire au minimum les superpositions.

figure 3 pour le plan easst lorsque $n = 50$ et 100 et pour le plan ppt lorsque $n = 50$ et 100. Nous avons tracé le graphique pour un sous-ensemble d'estimateurs de la variance. Dans chaque panneau, l'axe horizontal donne les valeurs de D^x . Puisque $v_j, v_j^{1/2}, v_j^{1/2}$ et $v_{R2}^{1/2}$ sont similaires pour la plupart de l'étendue de D^x , nous n'avons tracé que les points obtenus pour l'estimateur jackknife v_j . En outre, v_j^{DP} et v_j^{JP} étant proches, seul le dernier est représenté. L'estimateur GREG est entaché d'un biais conditionnel qui influe sur la regm si les échantillons ne sont pas équilibrés. Les propriétés conditionnelles médiocres de v_n sont surtout évidentes dans le cas de l'échantillonnage aléatoire simple pour lequel le biais de v_n en tant qu'estimation de l'eqm passe d'une valeur négative à une valeur positive sur l'étendue de D^x . Pour les autres estimations de la variance, le biais conditionnel est comparable au biais inconditionnel des deux v_{JP} et v_{SSW} sont tous deux approximativement dépourvus de biais par rapport au plan de sondage et au modèle et suivent bien la regm.

estimeurs incluant une correction pour la population finie fpc , c'est-à-dire v^*_{VSSW} , $v^{dp}_{V^*}$ et $v^*_{V^*}$, ont un taux de couverture plus proche du taux nominal de 95 %, tandis que $v^{r2}_{V^*}$, $v^{dp}_{V^*}$ et v_j couvrent environ 97 % ou 98 % des échantillons. L'estimateur $v^{fp}_{V^*}$, qui est une approximation du jackknife, mais qui n'inclut pas de fpc , est un bon choix indépendamment de la taille d'échantillon ou du plan

d'échantillonnage.

Les tableaux 4 et 5 montrent les taux de couverture pour la population active et la population active modifiée ($PA(mod)$). Les estimateurs $v^{dp}_{V^*}$, $v^{fp}_{V^*}$ et v_j sont nettement meilleurs pour la population active que pour la population active modifiée pour $n = 50$, aussi bien dans le cas de l'échantillonnage ppi. En revanche, pour $n = 250$, les taux de couverture sont les mêmes pour tous les estimateurs. L'estimateur fondé purement sur le plan de sondage, $v^{r2}_{V^*}$, donne de mauvais résultats lorsque la taille d'échantillon est petite pour l'un et l'autre plans d'échantillonnage. Comme pour la population d'hôpitaux, $v^{fp}_{V^*}$ produit une couverture quasiment égale au taux nominal pour chaque taille d'échantillon dans le cas de la population active.

Tableau 4

Taux de couverture des intervalles de confiance à 95 % pour des simulations au moyen de la population active et de la population active modifiée $PA(mod)$ et neuf estimateurs de la variance. En tout, 3 000 échantillons aléatoires simples ont été tirés sans remise pour des tailles d'échantillon de 50 et 100. L représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - \bar{Y})/\sqrt{v_{12}} < -1,96$; M représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - \bar{Y})/\sqrt{v_{12}} \leq 1,96$; U représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - \bar{Y})/\sqrt{v_{12}} > 1,96$.

Population active											
L			n=50			n=100			n=250		
v^*	v^{r1}	v^{r2}	v^{dp}	v^{fp}	v_j	v^*	v^{r1}	v^{r2}	v^{dp}	v^{fp}	v_j
5,3	4,9	4,2	4,3	4,3	4,3	2,9	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,8	2,8
4,9	4,9	4,2	4,3	4,3	4,3	2,7	2,8	2,8	2,9	2,	

Dans le cas de la population active et de la population active modifiée $PA(mod)$, l'augmentation de la taille de l'échantillon fait diminuer le biais. Les estimateurs $v^{\pi, R1}$, $v^{\pi, R2}$ et v^{SSW} sont entachés d'un biais négatif qui a tendance à devenir moins important à mesure que la taille de l'échantillon augmente. L'estimateur jackknife v_j et ses variantes, v_j , v_{jp} , surestiment la variance, particulièrement pour $n=50$. Pour chaque taille d'échantillon, les estimateurs v^D et v^{DP} sont presque dépourvus de biais dans de plus grandes proportions que la plupart des autres estimateurs.

Tableau 3

Taux de couverture des intervalles de confiance à 95 % pour des simulations au moyen de la population d'hôpitaux et de neuf estimateurs de la variance. En tout, 3 000 échantillons aléatoires simples ont été tirés sans remise pour des tailles d'échantillon de 50 et 100. L représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - Y) / \sqrt{I2} < -1,96$; M représente le pourcentage d'échantillons pour lesquels $|\hat{Y}_G - Y| / \sqrt{I2} \leq 1,96$; U représente le pourcentage d'échantillons pour lesquels $(\hat{Y}_G - Y) / \sqrt{I2} < 1,96$.

	$n=50$			$n=100$		
	L	M	U	L	M	U

Echantillonnage aléatoire simple	v^{π}	3,1	92,1	4,8	2,6	93,6
	v^{R1}	3,1	91,0	4,7	4,8	89,8
	v^{R2}	3,3	92,5	4,2	2,8	94,0
	v^{SSW}	3,3	92,5	4,2	2,8	94,0
	$v^{\pi, R1}$	4,2	91,0	4,7	4,8	89,8
	$v^{\pi, R2}$	2,8	93,9	3,3	3,3	97,0
	v^{DP}	3,1	93,0	3,9	2,7	94,3
	v^D	2,4	94,6	3,0	1,2	97,3
	v_j	2,2	95,0	2,8	1,2	97,3
	v_{jp}	2,9	93,6	3,5	2,6	94,6
Echantillonnage avec probabilité proportionnelle à la taille	v_j	2,2	95,1	2,8	1,2	97,4
	v_{jp}	2,9	93,6	3,5	2,6	94,6
	v^{π}	2,9	93,9	3,2	2,6	94,6
	v^{R1}	4,1	92,0	3,9	5,0	89,3
	v^{R2}	2,9	94,2	2,9	2,6	94,8
	v^{SSW}	2,1	95,8	2,1	0,9	98,3
	v^{DP}	2,7	94,5	2,8	2,5	95,0
	v^D	1,9	96,3	1,9	0,9	98,4
	v_j	2,6	94,8	2,6	2,4	95,4
	v_{jp}	1,7	96,5	1,8	0,8	98,4

Le tableau 3 donne la couverture empirique des intervalles de confiance à 95 % sur les 3 000 échantillons pour chaque ensemble, dans le cas de la population d'hôpitaux. Les trois estimateurs de la variance contenant des ajustements pour les effets leviers, mais non la correction pour la population finie jpc , c'est-à-dire v^D , v_j et v_{jp} , produisent une valeur plus grande et, donc, un taux de couverture plus élevé que $v^{\pi, R2}$ et v^{SSW} . (1996) ont, eux aussi constaté que l'estimateur jackknife a tendance à produire des estimations plus grandes de la variance de l'estimateur GREG que les autres estimateurs de la variance. Cette propriété est avantageuse dans le cas de la plus petite taille d'échantillon, $n = 50$. Si $n = 100$ et que la fraction d'échantillonnage est importante, les

Tableau 1
Biais relatif et racine carrée de l'erreur quadratique moyenne (reqm) de l'estimateur π et de l'estimateur de régression généralisée pour diverses études de simulation chacune 3 000 échantillons

Échantillonnage aléatoire simple						
		$n=50$		$n=100$		$n=250$
		Hôpitaux		Population active		PA(mod.)
\hat{p}_x	Bias rel. (%)	0,2	-0,1	-0,6	0	0
regm		76,6	50,7	34,2	24,1	15,5
\hat{p}_c	Bias rel. (%)	0,2	0,2	0,1	0,1	0,2
regm		32,6	21,1	28,3	19,9	12,4
Échantillonnage avec probabilité proportionnelle à la taille						
\hat{p}_x	Bias rel. (%)	0,2	0,2	0,1	0,2	0,4
regm		36,0	21,1	28,3	19,9	12,4
\hat{p}_c	Bias rel. (%)	0,2	0,2	0,1	0,2	0,4
regm		36,0	21,1	28,3	19,9	12,4

Echantillonnage aléatoire simple	v^{π}	0,1	0,1	0,5	0	0	-0,1
	v^{R1}	37,6	24,4	28,2	20,3	12,6	80,6
	v^{R2}	0,1	0,1	0,10	0	-0,6	-0,7
	v^{SSW}	0,1	0,1	0,10	0	-0,6	-0,7
	v^{DP}	27,2	16,9	28,2	19,3	12,0	81,8
	v^D	0,1	0,1	0,10	0	-0,6	-0,7
	v_j	0,1	0,1	0,10	0	-0,6	-0,7
	v_{jp}	37,6	24,4	28,2	20,3	12,6	80,6
	$v^{\pi, R1}$	0,1	0,1	0,5	0	0	-0,1
	$v^{\pi, R2}$	0,1	0,1	0,5	0	0	-0,1
Echantillonnage avec probabilité proportionnelle à la taille	v_j	0,1	0,1	0,10	0	-0,6	-0,7
	v_{jp}	37,6	24,4	28,2	20,3	12,6	80,6
	v^{π}	0,1	0,1	0,5	0	0	-0,1
	v^{R1}	37,6	24,4	28,2	20,3	12,6	80,6
	v^{R2}	0,1	0,1	0,10	0	-0,6	-0,7
	v^{SSW}	0,1	0,1	0,10	0	-0,6	-0,7
	v^{DP}	27,2	16,9	28,2	19,3	12,0	81,8
	v^D	0,1	0,1	0,10	0	-0,6	-0,7
	v_j	0,1	0,1	0,10	0	-0,6	-0,7
	v_{jp}	37,6	24,4	28,2	20,3	12,6	80,6

Tableau 2
Biais relatif de neuf estimateurs de la variance pour l'estimateur de régression généralisée lors de diverses études de simulation comptant chacune 3 000 échantillons

études de simulation comptant chacune 3 000 échantillons						
	Hôpitaux		Population active		PA(mod.)	
	$n=50$	$n=100$	$n=50$	$n=100$	$n=50$	$n=250$
V^{π}	-8,6	-4,2	-18,1	-12,3	-7,5	-16,3
V^{R1}	-18,9	-27	-11,3	-9,9	-8	-9,6
V^{SSW}	-7,6	-3	-10,9	-9,1	-5,9	-9,3
V^{R2}	5,9	30,1	-10,5	-8,2	-3,7	-8,8
V^{dp}	-1,4	32,3	0,1	-3,8	-3,8	0,6
V^{π}	13	34	0,6	-2,9	-1,6	1,6
V^{R1}	0,8	1,3	1	1	1	1,3

Echantillonnage aléatoire	v^{π}	-5,9	-0,9	-22,1	-12,1	-6,8	-16,5	-10,6	-0,3
	v^{R1}	-19,7	-32,4	-11,9	-7,7	-7,1	-9,1	-8,2	-2,7
	v^{R2}	-4	0	-11,6	-7	-4,9	-8,7	-7,3	-0,1
	v^{SSW}	16	52,6	-11,2	-6	-2,5	-8,3	-6,3	2,6
	v^{DP}	0,1	2	0,8	-0,3	-1,6	0,9	-2,5	2,1
	v^D	20,8	55,6	1,3	0,7	0,8	1,4	-1,5	4,8
	v_j	23,6	57,2	22,6	11,8	5,3	14,6	4,7	7,3
	v_{jp}	4,4	4	19,7	9,3	3,1	14,8	3,9	4,9
	$v^{\pi, R1}$	26,1	58,8	20,3	10,3	5,5	15,4	5	7,7
	$v^{\pi, R2}$	26,1	58,8	20,3	10,3	5,5	15,4	5	7,7
Echantillonnage avec probabilité proportionnelle à la taille	v^{π}	-5,9	-0,9	-22,1	-12,1	-6,8	-16,5	-10,6	-0,3
	v^{R1}	-19,7	-32,4	-11,9	-7,7	-7,1	-9,1	-8,2	-2,7
	v^{R2}	-4	0	-11,6	-7	-4,9	-8,7	-7,3	-0,1
	v^{SSW}	16	52,6	-11,2	-6	-2,5	-8,3	-6,3	2,6
	v^{DP}	0,1	2	0,8	-0,3	-1,6	0,9	-2,5	2,1
	v^D	20,8	55,6	1,3	0,7	0,8	1,4	-1,5	4,8
	v_j	23,6	57,2	22,6	11,8	5,3	14,6	4,7	7,3
	v_{jp}	4,4	4	19,7	9,3	3,1	14,8	3,9	4,9
	$v^{\pi, R1}$	26,1	58,8	20,3	10,3	5,5	15,4	5	7,7
	$v^{\pi, R2}$	26,1	58,8	20,3	10,3	5,5	15,4	5	7,7

La distribution résultante de $PA(mod)$ est présentée à la figure 1 où la rémunération hebdomadaire est représentée graphiquement en fonction du nombre d'heures travaillées pour les sous-groupes définis selon l'âge. Dans chaque panneau, les points noirs correspondent aux hommes et les cercles non remplis, aux femmes. Dans chaque panneau, une ligne de référence horizontale est tracée à 99\$. Bien qu'un nombre important de points se superposent, les caractéristiques générales sont évidentes. Le niveau de rémunération et la dispersion augmentent parallèlement à l'âge, le nombre d'heures travaillées par semaine est lié, quoiqu'assez faiblement, à la rémunération, et celle-ci présente l'asymétrie la plus forte pour le groupe des 25 à 34 ans et celui des 35 ans et plus. Le fait que la rémunération des hommes est généralement plus élevée que celle des femmes est moins évident.

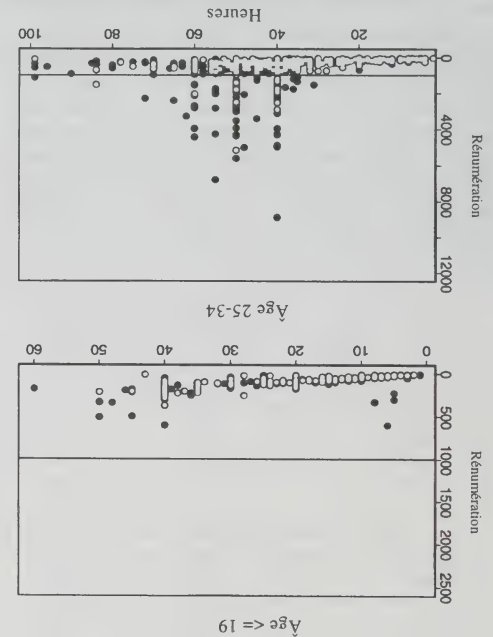


Figure 1.

Diagramme de dispersion de la rémunération hebdomadaire en fonction du nombre d'heures travaillées par semaine pour quatre groupes d'âge, pour la population $PA(mod)$. Les cercles vides correspondent aux femmes. Les cercles pleins correspondent aux hommes. Une ligne horizontale est tracée à 99\$, valeur maximale dans le cas de la population active originale.

La population d'hôpitaux compte $N = 393$ unités et une valeur auxiliaire unique x_i , qui est le nombre de lits dans chaque hôpital. La variable Y représente le nombre de patients ayant reçu leur congé durant une période de référence particulière. Pour cette population, l'estimateur GREG est fondé sur le modèle $E_M(X) = \beta_1 x^{1/2} + \beta_2 x$, $\text{var}_M(X) = \sigma^2 x$. Nous avons tiré des échantillons de taille égale à 50 et à 100 selon un plan d'échantillonnage aléatoire simple sans remise (cassr) et selon un plan d'échantillonnage avec probabilité proportionnelle à la taille (ppt) sans remise où la taille est égale à la racine carrée de x . Pour chaque combinaison de méthode d'échantillonnage et de taille d'échantillon, nous avons tiré 3 000 échantillons. Nous avons calculé les estimateurs $V_{R2}^1, V_{R2}^2, V_{R2}^3, V_{R2}^4, V_{R2}^5, V_{R2}^6, V_{R2}^7, V_{R2}^8, V_{R2}^9, V_{R2}^{10}, V_{R2}^{11}, V_{R2}^{12}, V_{R2}^{13}, V_{R2}^{14}, V_{R2}^{15}, V_{R2}^{16}, V_{R2}^{17}, V_{R2}^{18}, V_{R2}^{19}, V_{R2}^{20}, V_{R2}^{21}, V_{R2}^{22}, V_{R2}^{23}, V_{R2}^{24}, V_{R2}^{25}, V_{R2}^{26}, V_{R2}^{27}, V_{R2}^{28}, V_{R2}^{29}, V_{R2}^{30}, V_{R2}^{31}, V_{R2}^{32}, V_{R2}^{33}, V_{R2}^{34}, V_{R2}^{35}, V_{R2}^{36}, V_{R2}^{37}, V_{R2}^{38}, V_{R2}^{39}, V_{R2}^{40}, V_{R2}^{41}, V_{R2}^{42}, V_{R2}^{43}, V_{R2}^{44}, V_{R2}^{45}, V_{R2}^{46}, V_{R2}^{47}, V_{R2}^{48}, V_{R2}^{49}, V_{R2}^{50}, V_{R2}^{51}, V_{R2}^{52}, V_{R2}^{53}, V_{R2}^{54}, V_{R2}^{55}, V_{R2}^{56}, V_{R2}^{57}, V_{R2}^{58}, V_{R2}^{59}, V_{R2}^{60}, V_{R2}^{61}, V_{R2}^{62}, V_{R2}^{63}, V_{R2}^{64}, V_{R2}^{65}, V_{R2}^{66}, V_{R2}^{67}, V_{R2}^{68}, V_{R2}^{69}, V_{R2}^{70}, V_{R2}^{71}, V_{R2}^{72}, V_{R2}^{73}, V_{R2}^{74}, V_{R2}^{75}, V_{R2}^{76}, V_{R2}^{77}, V_{R2}^{78}, V_{R2}^{79}, V_{R2}^{80}, V_{R2}^{81}, V_{R2}^{82}, V_{R2}^{83}, V_{R2}^{84}, V_{R2}^{85}, V_{R2}^{86}, V_{R2}^{87}, V_{R2}^{88}, V_{R2}^{89}, V_{R2}^{90}, V_{R2}^{91}, V_{R2}^{92}, V_{R2}^{93}, V_{R2}^{94}, V_{R2}^{95}, V_{R2}^{96}, V_{R2}^{97}, V_{R2}^{98}, V_{R2}^{99}, V_{R2}^{100}$ pour chaque échantillon. Aux fins de comparaison, nous avons également inclus l'estimateur $\pi, \bar{Y} = \bar{Y}/N$. Nous avons inclus l'estimateur de la variance V_T , mais les résultats ne sont pas présentés ici parce qu'ils différaient fort peu de ceux obtenus pour V_{R2} .

La population active contient 10 841 personnes. Les variables auxiliaires que nous avons utilisées sont l'âge, le sexe et le nombre d'heures travaillées par semaine. La variable Y correspond à la rémunération hebdomadaire totale. Nous avons défini quatre groupes d'âge, à savoir 19 ans et moins, 20 à 24 ans, 25 à 34 ans et 35 ans et plus. Le modèle défini pour l'estimateur GREG inclut une ordonnée à l'origine, les effets principaux pour l'âge et le sexe et la variable quantitative correspondant au nombre d'heures travaillées. Nous avons posé que la variance du modèle était constante. Nous avons tiré des échantillons de taille égale à 50, 100 et 250, selon deux méthodes, à savoir l'échantillonnage sans remise avec probabilité proportionnelle au nombre d'heures travaillées. (Cette population présente certains regroupements par grappe, mais nous n'en avons pas tenu compte dans les simulations.)

La troisième population est une version modifiée de la population active conçue de façon à ajouter certaines observations aberrantes ou une certaine asymétrie dans les valeurs observées de la variable de rémunération hebdomadaire. Nous représentons cette nouvelle version par la notation « $PA(\text{mod})$ » aux fins de référence. Dans le cas de la population active originale, nous avons planifié la rémunération hebdomadaire à 99%. Pour chaque salaire planifié de la sorte, nous avons généré une nouvelle rémunération égale à 1 000% ainsi qu'une variable aléatoire logarithmique dont la distribution est caractérisée par des paramètres d'échelle et de forme égaux à 6,9 et 1, respectivement. Nous avons produit les rémunérations recodées pour 4,4 % de la population. Avant le recodage, la rémunération moyenne annualisée était de 19 359\$ et la valeur maximale, de 51 948\$; après le recodage, la moyenne était de 23 103\$ et la valeur maximale, de 608 116\$. Donc, $PA(\text{mod})$ est caractérisée par une courbe des revenus plus asymétrique que celle que l'on observerait pour la population réelle.

couverture des intervalles de confiance fondés sur l'estimateur jackknife est plus élevée que celui des intervalles de confiance fondés sur V_{R2}^{SSW} , ou V_T . Notons aussi que, si l'échantillonnage se fait sans remise et que certaines probabilités de sélection de premier ou de deuxième ordre sont faibles, les estimateurs $V_{R2}^1, V_{R2}^2, V_{R2}^3$ et V_T donneront lieu à une surestimation de la variance due au plan de sondage ou de la variance due au modèle. Pour tenir compte des probabilités de sélection non négligeables, nous pouvons procéder à certains ajustements simples. Une version ajustée de $V_T^*(X)$, modélisée sur V_{SSW} , est donnée par

$$V_T^{*P}(\hat{Y}_G) = \frac{1}{N^2} \sum_{i=1}^N \frac{\pi_i^2 (1 - h_i^2)}{(1 - \pi_i) g_i^2 r_i^2}.$$

Cette expression est similaire à l'estimateur V_{R2}^{JK3} de Duchesne (2000), quoique celui-ci ne tienne pas compte des effets leviers. L'expression (3.6) fait aussi penser à un autre estimateur qui est étroitement lié à un estimateur de la variance de l'erreur du meilleur prédicteur linéaire non biaisé de la moyenne dans les conditions du modèle (1.1) (voir Vaillant, Dorfman, et Royall 2000, chapitre 5). Cet estimateur est un peu moins conservateur que (3.6), mais comporte encore des corrections tenant compte des effets leviers :

$$V_D^P(\hat{Y}_G) = \frac{1}{N^2} \sum_{i=1}^N \frac{\pi_i^2 (1 - h_i^2)}{g_i^2 r_i^2}.$$

Comme $h_i = o(1)$, V_D est également approximativement dépourvu de biais par rapport au modèle et au plan de sondage. Une variante de cet estimateur qui pourrait donner de meilleurs résultats lorsque la valeur de certaines probabilités de sélection est grande est

$$V_D^{*P}(\hat{Y}_G) = \frac{1}{N^2} \sum_{i=1}^N \frac{\pi_i^2 (1 - h_i^2)}{(1 - \pi_i) g_i^2 r_i^2}.$$

4. RÉSULTATS DES SIMULATIONS

Pour déterminer la performance des estimateurs de la variance, nous avons réalisé plusieurs études de simulation en nous servant de trois populations différentes. La première est la population d'hôpitaux énumérée dans Vaillant, Dorfman et Royall (2000, Annexe B). La deuxième est la population active décrite dans Vaillant (1993). La troisième est une modification de la population active susmentionnée. Pour chacune de ces trois populations, l'échantillonnage a lieu sans remise, tel que décrit plus bas. Ces plans d'échantillonnage permettront de vérifier la notion selon laquelle les estimateurs de la variance motivés, en partie, par des plans d'échantillonnage avec remise demeurent utiles lorsqu'on les applique à des plans d'échantillonnage sans remise.

spécification plus générale, la variance approximative par rapport au modèle sous la spécification plus générale, $\text{var}^M(X^i) = \Psi^i$, est $(\Psi^i/n)(\bar{x}/\bar{x}^2)$ où $\Psi^i = \sum_s \bar{\Psi}^i/n$. La variance approximative par rapport au plan de sondage est $(1-f)/(nN) \sum_{i=1}^N (\bar{X}^i - \bar{x}_i)^2/\bar{x}^2$, où \bar{X}^i est une moyenne de population finie. L'estimateur $\bar{v}_{R2} = n^2(\bar{x}/\bar{x}^2)^2 \sum_s (\bar{X}^i - \bar{x}_i)^2/\bar{x}^2$ est approximativement sans biais pour la variance due au modèle et, parce que sans biais pour la variance due au plan de sondage, \bar{v}_{R2} est également approximativement sans biais pour la variance due au plan de sondage, à condition que f soit faible. En revanche, $\bar{v}_n = n^{-2}(1-f) \sum_s (\bar{X}^i - \bar{x}_i)^2/\bar{x}^2$ est approximativement sans biais par rapport au plan de sondage, mais n'est sans biais par rapport au modèle que pour des échantillons équilibrés où $\bar{x} = \bar{x}_s$. Royall et Cumberland (1981) ont observé des résultats comparables pour l'estimateur par ratio dans le cas de l'échantillonnage aléatoire simple sans remise.

3. AUTRES ESTIMATEURS DE LA VARIANCE BASÉS SUR LES CARRÉS CORRIGÉS DES RÉSIDUS

Le premier estimateur de rechange de la variance que nous considérons est le jackknife. La version particulière que nous étudions ici est définie comme étant

$$\bar{v}_J = \frac{n}{n-1} \sum_{i=1}^n [\bar{Y}^{G(i)} - \bar{Y}^{G(i)}]^2 \quad (3.1)$$

où $\bar{Y}^{G(i)}$ a la même forme que l'estimateur basé sur l'échantillon complet après omission de l'unité d'échantillonnage i . Si la probabilité de sélection a la forme $\pi_i = n\pi_i$, alors l'équation (3.1) peut être réécrite. Si nous adoptions comme convention de notation d'indiquer au moyen de l'indice (i) que l'unité d'échantillonnage i a été omise, nous avons

$$\begin{aligned} \bar{Y}^{G(i)} &= \bar{Y}^{G(i)}/N, \bar{Y}^{G(i)} = \sum_{s \neq i} \bar{Y}^{G(i)}/n, \bar{Y}^{G(i)} = \bar{\mathbf{T}}^{G(i)} + \bar{\mathbf{B}}^{(i)}(\mathbf{T}^x - \bar{\mathbf{T}}^{x(i)}), \\ \bar{F}^{\pi(i)} &= n \sum_{j=1}^f Y_j^f / [\pi_j(n-1)], \bar{F}^{x(i)} = n \sum_{j=1}^f x_j^f / [\pi_j(n-1)], \text{ et} \\ \bar{B}^{(i)} &= \bar{A}^{-1} \bar{X}^{s(i)} / \bar{V}^{-1} \Pi^{-1} \bar{Y}^{s(i)} \text{ où} \\ \bar{A}^{ss(i)} &= \bar{X}^{s(i)} \bar{V}^{-1} \Pi^{-1} \bar{X}^{s(i)} \end{aligned}$$

Une autre version, plus prudente, mais asymptotiquement équivalente, du jackknife consiste à remplacer $\bar{Y}^{G(i)}$ par l'estimateur basé sur l'échantillon complet $\bar{Y}^{G(i)}$. Les propriétés basées sur le plan de sondage de l'estimateur

incorrecte. Par conséquent, r_i^2 est un estimateur robuste de la forme de Ψ^i . Un estimateur simple, robuste, de la variance due au modèle pour l'unité i indépendamment de la forme de Ψ^i . Un estimateur simple, robuste, de la variance approximative due au modèle (1.5) est alors

$$\bar{v}_{R1}(\bar{Y}^{G(i)}) = N^{-2} \sum_s a_i^2 r_i^2 \quad (2.3)$$

qui est un genre d'estimateur « sandwich » (voir, par exemple, White 1982). (Notons qu'un argument formel pour montrer que \bar{v}_{R1} est robuste nécessiterait des conditions telles que $n^{-1}E_{M(i)}(r_{R1}^2)$ et $n^{-1}N^{-2} \sum_s a_i^2 \Psi^i$ convergent vers la même quantité.) Un autre estimateur de la variance, similaire à \bar{v}_{R1} si $\mathbf{a}_s = \Pi^T \mathbf{g}_s$, est

$$\bar{v}_{R2}(\bar{Y}^{G(i)}) = N^{-2} \sum_s \frac{\pi_i^2}{g_i^2} r_i^2. \quad (2.4)$$

Un estimateur de la variance approximative due au plan de sondage figurant dans (1.7) est

$$\bar{v}_n(\bar{Y}^{G(i)}) = N^{-2} \sum_s \frac{\pi_i^2}{1 - \pi_i} r_i^2. \quad (2.5)$$

Une autre solution, qui, selon Sæmål et coll. (1989), a de meilleures propriétés conditionnelles, est

$$\bar{v}^{SSW}(\bar{Y}^{G(i)}) = N^{-2} \sum_s \frac{1 - \pi_i}{\pi_i} \frac{\pi_i^2}{g_i^2} r_i^2. \quad (2.6)$$

Un autre estimateur, comparable, utilisé dans le logiciel SUPERCARP (Hidiroglou, Fuller et Hickman 1980) et obtenu par des méthodes de développement en série de Taylor, est

$$\bar{v}_T(\bar{Y}^{G(i)}) = N^{-2} \sum_s \left(\frac{\pi_i}{g_i r_i} - \frac{1}{n} \sum_s \frac{\pi_i^2}{g_i^2 r_i^2} \right)^2. \quad (2.7)$$

Comme nous le montrons à l'annexe, le deuxième terme entre parenthèses dans (2.7) converge en probabilité vers zéro dans les conditions du modèle (1.1). Donc, $\bar{v}_T \approx \bar{v}_{R2}$ dans le cas de grands échantillons.

Si la probabilité de sélection de chaque unité est faible, \bar{v}^{SSW} sera similaire à \bar{v}_{R1} , \bar{v}_{R2} et \bar{v}_T . Les trois estimateurs seront approximativement sans biais par rapport au plan de sondage dans les conditions de l'échantillonnage de Bernoulli et de Poisson. Par ailleurs, \bar{v}_n est approximativement sans biais par rapport au plan de sondage, mais ne tient pas compte des coefficients g_i et est biaisé en ce qui concerne tant le modèle (1.1) que le modèle (1.4).

À titre d'exemple simple, considérons l'échantillonnage de Bernoulli où $\pi_i = n/N$ et le modèle de travail $E_{M(i)}(X^i) = x_i^i \beta$, $\text{var}^M(X^i) = x_i^i x_i^i$. Alors, l'estimateur GREG est l'estimateur par ratio $\bar{Y}^{G(i)} = \bar{Y}_s^s \bar{x}^s / \bar{x}^s$ où \bar{x}^s est une moyenne de population finie. Dans le cas d'une

section 3.5, pour une description plus détaillée). L'échantillonnage de Bernoulli est un cas particulier de l'échantillonnage de Poisson où la probabilité d'inclusion est la même pour toutes les unités. En vertu de ces plans, la variance approximative due au plan de sondage de Y_G est

$$\text{var}^{\pi} (Y_G) = N^{-2} \sum_{i=1}^I \frac{\pi_i}{1 - \pi_i} E_i^2 \quad (1.7)$$

où $E_i = Y_i - x_i' B$ et $B = (X' V^{-1} X)^{-1} X' V^{-1} Y$ est

l'estimateur des paramètres de régression évalué pour la population finie complète. Särndal (1996) recommande d'utiliser l'estimateur GREG conjugué à des plans d'échantillonnage pour lesquels (1.7) est valide étant donné que la variance (1.7) est simple et que l'utilisation de l'estimation par régression peut souvent plus que compenser l'effet des tailles aléatoires d'échantillon qui sont une conséquence de ce genre de plan d'échantillonnage.

Les plans d'échantillonnage de Bernoulli et de Poisson, ainsi que les modèles linéaires (1.1) et (1.4) servent principalement de motif à l'utilisation des estimateurs de la variance présentés aux sections 2 et 3. Comme l'ont fait remarquer Yung et Rao (1996, page 24), il est courant d'utiliser des estimateurs de la variance qui sont appropriés pour un plan de sondage à tirages indépendants ou pour un plan de sondage avec remise, même si l'échantillon a été sélectionné sans remise. Parallelement, les estimateurs de la variance motivés par un modèle linéaire sont souvent appliqués à des cas où l'on s'attend à des divergences par rapport au modèle. Cette démarche pratique, qui sous-tend la logique suivie dans le présent article, est illustrée grâce à l'étude de simulation présentée à la section 4.

2. ESTIMATEURS DE LA VARIANCE

Dans le contexte de l'estimation de la variance, notre objectif général consistera à trouver des estimateurs convergents et approximativement non biaisés au rapport à un modèle que par rapport à un plan de sondage. Kot (1990) envisage aussi ce problème. Notons que l'objectif ici n'est pas d'estimer la variance combinée (ou anticipée) modèle-plan de sondage,

$$E^{\pi} E^{\pi} \left[\left(\bar{Y}_G - Y \right) - \left(E_M E^{\pi} (\bar{Y}_G - Y) \right) \right]^2.$$

Nous cherchons plutôt des estimateurs utiles à la fois pour $\text{var}^M (\bar{Y} - Y)$ et $\text{var}^{\pi} (\bar{Y})$. Les arguments utilisés comme en grande partie des arguments heuristiques comme justification des formes adoptées pour les estimateurs de la variance. En outre, nous devons appliquer des conditions formelles, telles que celles énoncées dans Royall et Cumberland (1978) ou Yung et Rao (2000) afin d'assurer la convergence et l'absence approximative de biais par rapport au modèle et au plan de sondage.

En premier lieu, considérons l'estimation de la variance approximative due au modèle donnée par (1.5). Dans le

développement qui suit, nous supposons que, à mesure que les valeurs de N et n deviennent grandes,

- i) $N \pi_j = O(n)$ et
- ii) \bar{A}^{π} / N converge vers une matrice de constantes

Le résidu associé à l'unité d'échantillonnage i est $r_i = Y_i - Y_j$ où $Y_j = x_j' B$. Le vecteur des valeurs prévues pour les unités d'échantillonnage peut s'écrire sous la forme

$$\bar{Y}^s = H^s Y^s \quad (2.1)$$

où $H^s = X^s A^{\pi s} V^{-1} X^s V^{-1} \Pi^{-1}$. La valeur prévue pour une unité individuelle est $Y_i^s = \sum_{j \in s} h_{ij}^s x_j^s$ où $h_{ij}^s = x_j^s A^{\pi s} V^{-1} \Pi^{-1} (v_j^s \pi_j^s)$ est le (ij) -ième élément de H^s . La matrice H^s est l'analogue de la matrice chapeau habituelle (Belsley, Kuh et Welsch 1980) de l'analyse par régression classique. Les éléments diagonaux de la matrice chapeau portent le nom d'effet levier et représentent une mesure de l'influence qu'exerce une unité sur sa propre valeur prévue. Notons que (2.1) prend en compte les inverses des probabilités de sélection, alors que ces derniers ne joueraient aucun rôle dans une analyse basée purement sur un modèle.

Le lemme qui suit, qui est une variante de certains résultats du Lemme 5.3.1 de Valliant et coll. (2000), donne certaines propriétés des effets leviers et de la matrice chapeau.

Lemme 1. Supposons que (i) et (ii) sont vérifiées. Pour $H^s = X^s A^{\pi s} V^{-1} \Pi^{-1}$, les propriétés qui suivent sont vérifiées pour tous les s :

- a) $h_{ij}^s = O(n^{-1})$;
- b) H^s est idempotente;
- c) $0 \leq h_{ii}^s \leq 1$.

Preuve : Puisque $h_{ij}^s = x_j^s A^{\pi s} V^{-1} \Pi^{-1} (v_j^s \pi_j^s)$, les conditions (i) impliquent que $h_{ij}^s = O(n^{-1})$. La partie (b) découle de la multiplication directe, en utilisant la définition de H^s . Pour prouver (c), notons que $h_{ii}^s \geq 0$, puis qu'il s'agit d'une forme quadratique. La partie (b) implique que $h_{ii}^s = h_{ii}^s + \sum_{j \neq i} h_{ij}^s h_{ji}^s$ qui est vérifiée uniquement si $h_{ii}^s \leq 1$. Ensuite, nous écrivons le résidu sous la forme $r_i = Y_i (1 - h_{ii}^s) - \sum_{j \in s(i)} h_{ij}^s Y_j$ où $s(i)$ est l'échantillon dont est exclue l'unité i . Puisque $E_M(r_i) = 0$, nous avons $E_M(r_i^2) = \text{var}^M(r_i)$ et

$$E_M(r_i^2) = \Psi_i (1 - h_{ii}^s)^2 + \sum_{j \in s(i)} h_{ij}^s \Psi_j \quad (2.2)$$

dans les conditions du modèle (1.4). En utilisant le lemme 1(a), nous avons $h_{ii}^s = o(1)$, $h_{ij}^s = o(1)$ et, conséquemment, $E_M(r_i^2) = \Psi_i$. Donc, dans le cas de grands échantillons, r_i^2 est un estimateur approximativement non biaisé de la variance due au modèle correct, même si la spécification de la variance dans le modèle (1.1) est

donné x par la distribution des indicateurs étant donné x dans ce cas, l'inférence fondée sur un modèle peut se faire (voir Sugden et Smith 1984 pour une définition formelle).

sélection.

Le vecteur des n cibles pour les unités d'échantillonnage

est $Y = (Y_1, \dots, Y_n)'$ et la matrice $n \times p$ des variables auxiliaires pour les unités d'échantillonnage est

$X_s = (x_1, \dots, x_n)'$. Définissons la matrice diagonale des probabilités de sélection comme étant $\Pi_s = \text{diag}(\pi_1, \dots, \pi_n)$, et la matrice diagonale des variances dues au modèle, comme étant $V_s = \text{diag}(v_1, \dots, v_n)$. L'estimateur GREG du total,

$T = \sum_{i=1}^N Y_i$, est alors défini comme étant l'estimateur d'Horvitz-Thompson ou l'estimateur π , $\pi = \sum_{i=1}^N X_i / \pi_i$, plus un ajustement :

$$(1.2) \quad \hat{T}_G = \hat{T}_\pi + \hat{B}'(T_\pi - \hat{T}_\pi)$$

où $\hat{B} = A_{\pi_s}^{-1} X_s' V_s^{-1} \Pi_s^{-1} Y_s$ avec $A_{\pi_s} = X_s' V_s^{-1} \Pi_s^{-1} X_s$, et $\hat{T}_\pi = \sum_{i=1}^N x_i / \pi_i$. Nous pouvons aussi écrire l'estimateur GREG sous la forme

$$(1.3) \quad \hat{T}_G = g_s' \Pi_s^{-1} Y_s$$

où $g_s = V_s^{-1} X_s' A_{\pi_s}^{-1} (T_\pi - \hat{T}_\pi) + 1$ et 1 est un vecteur de n valeurs. L'expression (1.3) sera utile pour les calculs

subsequents.

Une variante de l'estimateur GREG, appelée estimateur

« esthétique », a été proposée par Särndal et Wright (1984) et amplifiée par Brewer (1995, 1999). Un estimateur

de la variance présente ici pourrait également être adaptés de façon à couvrir l'estimation esthétique.

A condition que l'on connaisse N , l'estimateur GREG de la moyenne est simplement $\hat{Y}_G = \hat{T}_G / N$. Nous nous

concentrerons ici sur l'analyse de Y_G . (Dans certaines situations, particulièrement celles où l'on recourt à l'échantillonnage à plusieurs degrés, on ne connaît pas la taille de la population et il faut utiliser une estimation, N , dans le

dénominateur de Y_G . L'analyse qui suit concernant la moyenne ne s'applique pas dans ce cas.) Nous pouvons

utiliser dans l'estimateur GREG des variables auxiliaires quantitatives ou qualitatives (ou les deux). Si nous utilisons

une variable qualitative, comme le sexe (masculin ou féminin), alors deux colonnes ou plus de la matrice X_s

seront linéairement dépendantes, auquel cas nous utiliserons dans (1.2) et (1.3) une inverse généralisée,

représentée par $A_{\pi_s}^{-}$. Notons que, même si A_{π_s} n'est pas unique, l'estimateur GREG ne varie pas en fonction du

choix de l'inverse généralisée. La preuve est similaire au théorème 7.4.1 dans Valliant et coll. (2000).

L'estimateur GREG est non biaisé en ce qui concerne le modèle dans les conditions de (1.1) et est approximativement non biaisé en ce qui concerne le plan de sondage dans le cas de grands échantillons probabilistes. Notons que

$E_M(Y) = x_1' \beta$; si les paramètres de variance sont mal précisés dans (1.1), l'estimateur GREG demeurera non biaisé en ce qui concerne le modèle. Par contre, si $E_M(Y)$ est définie comme étant

$$\hat{Y}_G - \bar{Y} = N^{-1} (a_s' Y_s - 1' Y)$$

où $\bar{Y} = T/N$, $a_s = \Pi_s^{-1} g_s - 1$, X_s' est le vecteur $(N - n)$ des variables cibles pour les unités non échantillonnées, et 1 est un vecteur de $N - n$ valeurs. 1. Supposons maintenant que le modèle réel de Y_i est

$$E_M(Y_i) = x_i' \beta$$

$$(1.4) \quad \text{var}_M(Y_i) = \psi_i$$

autrement dit que la spécification de la variance est différente de celle donnée dans (1.1), mais que $E_M(Y_i)$ est la même. Si nous utilisons l'erreur d'estimation, la variance de l'erreur de Y_G est alors

$$\text{var}_M(Y_G - \bar{Y}) = N^{-2} (a_s' \Psi_s a_s + 1' \Psi_s 1)$$

où la matrice des covariances $n \times n$ de Y_s est $\Psi_s = \text{diag}(\psi_1)$ et Ψ_s est la matrice des covariances $(N - n) \times (N - n)$ de Y . Si les tailles d'échantillon et de

population sont toutes les deux importantes et que la fraction d'échantillonnage, $f = n/N$, est négligeable, la

variance de l'erreur est approximativement

$$(1.5) \quad \text{var}_M(Y_G - \bar{Y}) \approx N^{-2} \sum_{i=1}^n a_i' \psi_i a_i$$

Notons que cette variance dépend des paramètres de la variance réelle, ψ_i , et des paramètres de la variance du modèle de travail, v_i , car v_i fait partie de a_i . Puisque a_s est approximativement identique à $\Pi_s^{-1} g_s$ lorsque les probabilités de sélection sont faibles, la variance de l'erreur

est, dans ce cas, aussi approximativement égale à

$$(1.6) \quad \text{var}_M(Y_G - \bar{Y}) \approx N^{-2} \sum_{i=1}^n \frac{\pi_i^2}{2} \psi_i$$

En ce qui concerne l'estimation de la variance due au modèle, nous prendrions comme cible l'une ou l'autre des formes asymptotiques (1.5) et (1.6). Cependant, si la fraction d'échantillonnage est grande, le terme $1' \Psi_s 1 / N^2$ peut représenter une part importante de la variance de l'erreur et (1.5) ou (1.6) pourrait constituer une

mauvaise approximation.

Nous considérerons la variance due au plan de sondage dans le cas des plans d'échantillonnage à un degré de

Bernoulli et de Poisson. Dans le cas de l'échantillonnage de Poisson, les indicateurs δ_i de la présence ou non d'une

unité dans l'échantillon sont indépendants et $P(\delta_i = 1) = 1 - P(\delta_i = 0) = \pi_i$ (voir Särndal, Swensson et Wretman 1992,

Estimation de la variance de l'estimateur de régression généralisée

RICHARD VALLANT¹

RÉSUMÉ

On trouve dans la littérature sur l'échantillonnage diverses propositions d'estimateurs de la variance de l'estimateur de régression généralisée (GREG) d'un moyen, principalement en vue d'estimer la variance due au plan de sondage. Il est facile de concevoir des estimateurs de la variance qui, dans certaines conditions, sont approximativement non biaisés en ce qui concerne le plan de sondage et le modèle. Nous étudions ici plusieurs estimateurs bivariés dans le cas de l'échantillonnage à un degré. Il s'agit d'estimateurs robustes de la variance due au modèle, même si le modèle qui motive l'estimateur de régression généralisée (GREG) comprend un paramètre de variance incorrect. L'une des caractéristiques principales des estimateurs robustes est le rajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Nous montrons aussi que l'estimateur jackknife avec suppression d'une unité inclut les ajustements pour tenir compte des effets de leviers et est un bon choix du point de vue tant de la variance due au plan de sondage que de celle due au modèle. Dans un ensemble de simulations, ces estimateurs de la variance sont caractérisés par un biais faible et produisent des intervalles de confiance dont le taux de couverture est quasi nominal pour plusieurs méthodes d'échantillonnage, tailles d'échantillon et populations dans le cas de l'échantillonnage à un degré.

Nous présentons aussi les résultats des simulations pour une population à distribution asymétrique où tous les estimateurs de la variance donnent de mauvais résultats. Les échantillons qui ne représentent pas adéquatement les unités dont la valeur est grande produisent des estimations de la moyenne trop faibles, des estimations de la variance trop faibles et des intervalles de confiance dont la couverture est nettement inférieure au taux nominal. Ces faiblesses doivent être évitées à l'étape de l'élaboration du plan de sondage grâce à la sélection d'échantillons qui couvrent bien les unités extrêmes. Cependant, ceci n'est pas faisable dans le cas de populations pour lesquelles les renseignements sur le plan de sondage sont insuffisants.

MOTS CLÉS : Couverture de l'intervalle de confiance; matrice chapeau; jackknife; effet levier; modèle non biaisé; asymétrie.

1. INTRODUCTION

L'estimation robuste de la variance est un élément important dont il faut tenir compte en cas d'échantillonnage d'une population finie où on utilise l'approche prédictive. Valliant, Dorfman et Royall (2000) résument une grande partie des études publiées dans le domaine de la modélisation. Selon l'approche prédictive, on formule un modèle de travail que l'on utilise ensuite pour construire un estimateur ponctuel d'une moyenne ou d'un total. On crée des estimateurs de la variance qui sont robustes en ce sens qu'ils sont approximativement sans biais par rapport au modèle et convergent pour la variance due au modèle, même si la spécification de la variance dans le modèle de travail est incorrecte. Dans le présent article, nous étendons cette approche à l'estimateur de régression généralisée (GREG) afin de construire des estimateurs de la variance qui sont approximativement non biaisés en ce qui concerne le modèle, mais qui sont aussi approximativement non biaisés en ce qui concerne le plan de sondage dans le cas de l'échantillonnage à un degré. Nous comparons plusieurs solutions, y compris l'estimateur jackknife et certaines variantes de ce dernier. Nous utilisons une classe particulière de modèles linéaires ainsi qu'un échantillonnage de Bernoulli ou de Poisson comme fondement des estimateurs de la variance. Cependant, en pratique,

certaines de ces estimateurs peuvent être appliquées avec de bons résultats à des plans d'échantillonnage à un degré où les tirages ne sont pas indépendants. À chaque unité de la population sont associées une variable cible y_i et un vecteur p de variables auxiliaires $x_i = (x_{i1}, \dots, x_{ip})'$, où $i = 1, \dots, N$. Le vecteur des totaux de la population des variables auxiliaires est $T_x = (T_{x1}, \dots, T_{xp})'$ où $T_{xk} = \sum_{i=1}^N x_{ik}$, $k = 1, \dots, p$. L'estimateur de régression généralisée, défini plus bas, est motivé par un modèle linéaire dans lequel les X sont des variables aléatoires indépendantes telles que

$$E_M(Y_i) = x_i' \beta$$
$$\text{var}_M(Y_i) = v_i.$$

(1.1)

Dans la plupart des situations, (1.1) est un modèle « de travail » vraisemblablement incorrect dans une certaine mesure. Supposons que l'on sélectionne un échantillon probabiliste s et que la probabilité de sélection de l'unité d'échantillonnage i soit $P(i) = \pi_i$, où δ_i est un indicateur 0-1 de la présence ou non de l'unité dans l'échantillon. Nous supposons que le mécanisme de sélection de l'échantillon est ignorable. Approchativement parlant, un mécanisme ignorable signifie que la distribution conjointe des X et des indicateurs d'échantillonnage, étant donné les x_i , peut être décomposée en un produit de la distribution de X étant

¹ Richard Valliant, Westat, 1650 Research Boulevard, Rockville, MD 20850.

BIBLIOGRAPHIE

BRICK, J.M., et MORGANSTEIN, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.

BRICK, J.M., et MORGANSTEIN, D. (1997). Computing sampling errors from clustered unequally weighted data using replication: WesVarPC. *Bulletin of the International Statistical Institute. Proceedings*, 1, 479-482.

COX, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 721-726.

DIPPO, C.S., FAY, R.E. et MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 489-494.

KOTT, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et SHAO, J. (1996). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Society*, 91, 343-348.

RAO, J.N.K., et SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RUST, K., et RAO, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381-397.

SHAO, J. (1996). Resampling methods in sample surveys (avec discussion). *Statistics*, 27, 203-254.

SHAO, J., et CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhyā*, B, Special Issue on Sample Surveys, 187-201.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Society*, 93, 819-831.

SHAO, J., et STEEL, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Society*, 94, 254-265.

SHAO, J., et TU, D. (1995). *The jackknife and Bootstrap*. New York: Springer-Verlag.

VALLANT, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

logiciel afin d'utiliser les $\hat{y}_{hij}^{(v)}$ plutôt que $\hat{y}_{hij}^{(v)}$ lors du calcul des estimations répétées. Les principaux inconvénients tiennent au fait que 1) il faudrait procéder à une reprogrammation complexe du logiciel, 2) s'il faut imputer les valeurs de plusieurs variables, le nombre de champs de données nécessaires augmente considérablement et 3) la méthode selon laquelle un analyste des données estimerait la variance d'une variable dérivée, disons d , n'est pas certaine, à moins que les $d_{hij}^{(v)}$ figurent d'avance dans le fichier. Les caractéristiques avantageuses de cette mise en oeuvre sont que 1) aucun entreregistrement supplémentaire n'est nécessaire et 2) l'estimation de la variance pour des sous-domaines ne nécessite pas de travaux supplémentaires.

7. CONCLUSIONS

Les méthodes d'estimation de la variance par rééchantillonnage corrigé de Rao et Shao (1992) et de Shao, Chen et Chen (1998) offrent un moyen de calculer des estimations de la variance qui tiennent compte de l'imputation pour compenser la non-réponse partielle. Une future étape importante consistera à trouver des moyens de faciliter les calculs. Le présent article décrit les applications fondées sur l'utilisation de poids de rééchantillonnage.

REMERCIEMENTS

Une question posée par Robert E. Fay lors d'une présentation à la Washington Statistical Society a fourni l'idée à l'origine du présent article. L'auteur remercie également les deux évaluateurs, le rédacteur en chef adjoint et le rédacteur en chef de leurs commentaires utiles. L'auteur travaillait pour le National Center for Education Statistics au moment de la rédaction de la version initiale de l'article. L'article expose les opinions personnelles de l'auteur et ne sous-entend pas que ces opinions bénéficient de l'appui officiel du U.S. Department of Education ou du U.S. Department of Transportation ni ne permet d'intéresser un tel soutien.

Souignons que Rao et Shao (1992) n'ont proposé et évalué leur méthode d'estimation jackknife de la variance que pour les estimations des totaux (ou des moyennes). Il convient donc de se montrer prudents lors de l'utilisation de la méthode pour des statistiques plus complexes. Pareillement, Shao, Chen et Chen (1998) ont proposé leur méthode d'estimation de la variance par rééchantillonnage équilibré pour des fonctions des totaux et pour des quantiles, si bien qu'il ne faut pas l'appliquer à d'autres statistiques.

5. MÉTHODE HOT DECK PONDÉRÉE

Le recours à l'imputation par la méthode hot deck pondérée (par exemple Cox 1980) présente plusieurs avantages, si bien que nous y consacrons une section distincte. Rao et Shao (1992) se concentrent sur cette méthode d'imputation qui est également discutée dans Shao, Chen et Chen (1998). En vertu de cette méthode, pour remplacer une réponse manquante, on impute une valeur sélectionnée au hasard à partir des réponses à la question étudiée qui figurent dans la classe d'imputation. La probabilité de sélection est proportionnelle à $w_{h'f'}$, c'est-à-dire le poids appliqué au répondant. Les répondants pour lesquels la probabilité d'être sélectionné est positive sont appelés des *donneurs potentiels*; le non-répondant qui est visé par l'imputation est le receveur. S'il existe dans le fichier plus d'une question pour lesquelles on imputera une valeur par la méthode hot deck pondérée, la situation est simplifiée si l'on utilise comme donneurs potentiels des répondants ayant fourni une réponse complète (unités qui ont répondu à toutes les questions) et que l'on n'utilise qu'un seul donneur pour imputer les valeurs pour toutes les questions d'un receveur particulier nécessitant une imputation par la méthode hot deck pondérée. (Le donneur est sélectionné par la méthode hot deck pondérée, la situation est la même si la probabilité de répondre à une question est la même pour chaque unité d'une classe d'imputation, la méthode hot deck pondérée produit des estimations des moyennes, de totaux et des quantiles d'échantillon conformes au plan de sondage. De surcroît, les valeurs imputées sont « plausibles » en ce sens qu'elles ont l'air de données réelles.

Une caractéristique intéressante de la méthode hot deck pondérée est son équivalence en cas de transformations objectives. Pour expliquer l'équivalence, considérons la variable dérivée d , où $d = g(y)$ et g est une fonction bijective. Alors, par la méthode hot deck pondérée, nous répondons à la question au moyen de $y_{h'f'}$ et nous utilisons $g(y_{h'f'})$ pour d . Ceci revient à utiliser la méthode hot deck pondérée pour imputer la valeur de d au moyen de $d_{h'f'}$ et à utiliser $g^{-1}(d_{h'f'})$ pour y . Peu d'autres méthodes présentent cette caractéristique de l'imputation hot deck. Par exemple, en cas d'imputation par la moyenne

6. AUTRES MÉTHODES

pour toutes les variables.

$$w_{h'f'}^{(r)} = (a_{h'f'; h'f'; h'f'} - a_{h'f'; h'f'; h'f'}) w_{h'f'}^{(r)}$$

Supposons que l'on applique la méthode hot deck pondérée à plusieurs variables d'un fichier et que l'on se serve de répondants ayant fourni des réponses complètes comme donneurs potentiels. Dans ce cas, même si le profil de non-réponse diffère selon la variable à laquelle il faut imputer des valeurs, l'application du rééchantillonnage corrigé au moyen des poids de rééchantillonnage décrite à la section précédente peut se faire en utilisant le même ensemble de poids de rééchantillonnage supplémentaires

À la présente section, nous considérons d'autres méthodes, y compris une qui nécessite la modification du logiciel.

6.1 Première méthode de rechange

Un moyen de réduire le nombre d'enregistrements consiste à inclure des enregistrements supplémentaires de la forme

$$ID' \quad IC \quad 0 \quad w_{h'f'}^{(R)} \quad y_{h'f'} \quad IF' \quad 0 \quad IF'$$

où ID' est l'identificateur de l'unité donneuse potentielle (h', f') qui a répondu à la question y , B_k est l'ensemble des unités non répondantes à la question y dans la classe d'imputation k et

$$w_{h'f'}^{(r)} = \sum_{(h'f') \in B_k} (a_{h'f'; h'f'; h'f'} - a_{h'f'; h'f'; h'f'}) w_{h'f'}^{(r)}$$

$$r = 1, \dots, R.$$

Dans ces conditions, pour une question donnée, il n'existe qu'un seul enregistrement supplémentaire par donneur potentiel. Le principal inconvénient est que, à cause de la sommation, il est impossible de calculer les estimations pour des sous-domaines qui recoupent plusieurs classes d'imputation.

6.2 Deuxième méthode de rechange

L'application la plus évidente consisterait peut-être à ajouter les $y_{h'f'}$ à l'enregistrement $(h'f')$ et à modifier le

Illustration numérique : Partie du fichier de données utilisé pour l'estimation de la variance

ID	IC	$w_{hij}^{(1)}$	$w_{hij}^{(R)}$	\bar{y}_{hij}	IF_y	\bar{z}_{hij}	IF_z
001	1	10,1	20,2000	...	0,0000	5,4	1,2
002	1	20,3	40,6000	...	0,0000	5,1	0
003	1	18,4	36,8000	...	0,0000	5,2	0
004	1	11,1	0,0000	...	22,2000	5,1	1
005	1	16,3	0,0000	...	32,6000	5,1	1
006	1	15,4	0,0000	...	30,8000	5,4	0
001	1	0,0	3,0162	...	0,0000	5,1	2
001	1	0,0	2,7339	...	0,0000	5,2	2
001	1	0,0	-5,7501	...	0,0000	5,4	2
004	1	0,0	0,0000	...	-8,3301	5,1	2
004	1	0,0	0,0000	...	-7,5505	5,2	2
004	1	0,0	0,0000	...	-12,2325	5,4	2
005	1	0,0	0,0000	...	-11,0876	5,2	2
005	1	0,0	0,0000	...	23,3201	5,4	2
001	1	0,0	5,5645	...	0,0000	0,0	3
001	1	0,0	5,0436	...	0,0000	0,0	3
001	1	0,0	-2,7512	...	0,0000	0,0	3
001	1	0,0	-4,0400	...	0,0000	0,0	3
001	1	0,0	-3,8169	...	0,0000	0,0	3

Dans l'illustration numérique du tableau 1, les neuf enregistrements (lignes du tableau) pour lesquels $IF_y = 2$ sont les enregistrements supplémentaires pour la question y. Les six premiers enregistrements sont les enregistrements originiaux pour les six unités d'échantillonnage qui représentent la classe d'imputation $IC = 1$. (Les enregistrements figurant à la fin pour lesquels $IF_z = 2$, qui sont les enregistrements supplémentaires pour la question z, seront examinés à la fin du paragraphe. Dans ces enregistrements, le signal d'imputation pour y, IF_y , a été fixé à 3 pour indiquer qu'ils correspondent à des enregistrements supplémentaires pour une autre question que y.) Il existe trois répondants ($IF_y = 0$) et trois non-répondants ($IF_y = 1$) à la question y. Nous supposons que la méthode d'imputation est la méthode hot deck pondérée. Nous présentons uniquement les premier et dernier poids de rééchantillonnage ($w_{hij}^{(1)}$ et $w_{hij}^{(R)}$), mais ils concordent avec les poids de rééchantillonnage utilisés pour la méthode par rééchantillonnage équilibré d'estimation de la variance. Nous avons $\sum w_{hij}^{(1)} \bar{y}_{hij} = 476,650$, $\sum w_{hij}^{(1)} \bar{z}_{hij} = 506,048$ et $\sum w_{hij}^{(R)} \bar{y}_{hij} = 455,696$, où les sommes sont calculées sur l'ensemble des enregistrements. Le lecteur peut vérifier que ces résultats concordent avec $\sum w_{hij}^{(R)} \bar{y}_{hij} = 476,650$, $\sum w_{hij}^{(R)} \bar{z}_{hij} = 506,048$ et $\sum w_{hij}^{(1)} \bar{y}_{hij} = 506,048$ obtenus au moyen de (3.1), où les sommes sont calculées sur les six premiers enregistrements uniquement.

où les poids $w_{hij}^{(1)} = w_{hij}^{(1)} : h, i, j, \dots, w_{hij}^{(R)} = w_{hij}^{(R)} : h, i, j, \dots$ sont calculés selon (4.1), mais en utilisant la méthode d'imputation et le profil de réponse correspondant à la question z. La méthode d'imputation pour z ne doit pas nécessairement être la même que celle utilisée pour y, mais doit avoir la forme décrite à la section 3. Dans le tableau 1, les enregistrements supplémentaires pour la question z sont ceux pour lesquels $IF_z = 2$. Nous avons alors $\sum w_{hij}^{(1)} \bar{z}_{hij} = 120,30$, $\sum w_{hij}^{(1)} \bar{z}_{hij} = 124,349$ et $\sum w_{hij}^{(R)} \bar{z}_{hij} = 115,400$, où les sommes sont calculées sur l'ensemble des enregistrements. Ces résultats concordent avec les sommes obtenues selon l'équation (3.1).

Malheureusement, le plus gros inconvénient de cette méthode tient au grand nombre d'enregistrements supplémentaires qu'il faut ajouter au fichier. Cet inconvénient est moins prononcé lorsque les classes d'imputation sont petites. (Cependant, la taille des classes d'imputation dépend de nombreux facteurs.) Par contre, les avantages sont les suivants :

- Les estimations répétées et les estimations de la variance fondées sur les rééchantillonnages corrigés peuvent être calculées au moyen de n'importe quel logiciel conçu pour estimer la variance en se fondant sur les poids de rééchantillonnage.
- Si l'existe une autre variable, disons y', présentant le même profil de non-réponse que y et pour laquelle on utilise exactement la même méthode d'imputation que pour y (autrement dit, les mêmes valeurs de a et $a^{(y)}$), le calcul des estimations répétées pour y' peut être réalisé sans ajouter de nouveaux enregistrements.
- Des estimations peuvent être calculées sur des sous-domaines, même s'ils recoupent les limites des classes d'imputation.
- Si l'on suppose que la méthode d'imputation est la méthode hot deck pondérée, on estime la variance de la méthode hot deck pondérée, disons log y où $y > 0$, en ajoutant simplement la variable dérivée à chaque enregistrement et en se fondant sur cette variable pour calculer les estimations répétées. (Nous ferons d'autres commentaires sur la méthode hot deck pondérée à la section suivante.)

L'analyste des données peut choisir de supprimer les enregistrements supplémentaires d'une copie du fichier de données et utiliser le fichier réduit pour repérer les valeurs aberrantes, formuler des hypothèses, etc., puis réintégrer les enregistrements supplémentaires dans le fichier au moment d'estimer les variances.

Donc,

(h', i', j') représente une unité qui a répondu à la question Y .

$$E_{A_k}^{(r)}(\tilde{y}_{h_{j_0}}) = \sum_{(h', i', j') \in A_k} a_{(r)}^{(h', i', j'; h_{j_0}, i_{j_0}, j_{j_0})} y_{h', i', j'}$$

ou les $a_{i,j}, b_{i,j}$ et $a_{i,j}, b_{i,j}$ sont des constantes qui ne dépendent pas des valeurs de i, j et $a_{i,j}, b_{i,j} = 0$

3.1 Exemple : Imputation par quotient

à imputer une valeur pour une réponse manquante $Y_{h,i,j}^*$.

$$\sum_{\substack{f, l, l, y \in \mathcal{X} \\ f, l, l, y \in \mathcal{M}}} \frac{1}{\sum_{f, l, l, y \in \mathcal{M}} 1} = \frac{1}{\sum_{f, l, l, y \in \mathcal{M}} 1} \sum_{f, l, l, y \in \mathcal{M}} 1$$

Notons que les $a^{h_{i,j}; h_{i,j}^0}$ et $a^{h_{i,j}; h_{i,j}^0}$ dépendent des

Cette méthode d'imputation consiste à remplacer une

Chen et Chen (1998, page 822) montrent que

$$\sum_{(f_1, \dots, f_k) \in A^k} f_1 u_1 \otimes \dots \otimes f_k u_k = \sum_{(f_1, \dots, f_k) \in A^k} f_1 u_1 \otimes \dots \otimes f_k u_k$$

l'enregistrement. Un enregistrement ressemblera à :

L'ESTIMATION DE LA VARIANCE

$$\bigcup_{(h', i', j', u) \in A_k} f_1 i' u \quad / \quad f_1 i u \quad f_1 u' i' u$$

$${}^2H \ 0 \ \kappa \ H \ f_{1,1} \gamma \ f_{1,1} \gamma \cdot f_{1,1} \gamma \ M \dots f_{1,1} \gamma \cdot f_{1,1} \gamma \ M \ 0 \ 2H \ 0$$
$$r = 1, \dots, R. \quad (4.1)$$

peuvent être négatifs.

3. MÉTHODES PAR RÉÉCHANTILLONNAGE CORRIGÉ

Les travaux de Rao et Shao (1992) et de Shao, Chen et Chen (1998) servent de fondement au présent article. Shao et Chen (1999) et Shao et Steel (1999) traitent également de l'estimation de la variance des données d'enquête par rééchantillonnage en cas d'imputation.

Nous commençons par décrire la notation, en nous inspirant en grande partie de celle utilisée par Shao, Chen et Chen (1998). La population est divisée en L strates de sorte que N_h grappes soient comprises dans la $h^{ième}$ strate. À la première étape d'échantillonnage dans la strate h , nous tirons $m_h \geq 2$ grappes, la $i^{ième}$ grappe étant sélectionnée avec la probabilité $p_{hi}, i = 1, \dots, N_h; h = 1, \dots, L$. Les grappes sont sélectionnées sans remise et de façon indépendante dans les diverses strates. Nous supposons que les fractions d'échantillonnage n_h/N_h sont suffisamment faibles pour éviter d'apporter une correction pour les populations finies. D'autres étapes d'échantillonnage peuvent avoir lieu dans chaque grappe, de façon indépendante avoir lieu dans chaque grappe, de façon indépendante de grappe en grappe. La grappe i de la strate h contient, en dernière analyse, N_{hi} unités de population. Pour chaque unité de population (h, i, j) , il existe une variable étudiée y_{hij} . Posons que S représente l'ensemble des unités d'échantillonnage et que $\{y_{hij}, (h, i, j) \in S\}$ est l'ensemble des données qui a fait l'objet d'une imputation : les y_{hij} sont égales aux y_{hij} lorsque la variable est observée et égales à la valeur imputée autrement. Les unités notées au moyen de l'indice k et A_k représentent l'ensemble des répondants pour la variable y dans la classe d'imputation k . Nous supposons que l'ensemble de données contient des identificateurs (« signaux ») qui permettent de repérer les non-répondants.

Dans le cas des méthodes par rééchantillonnage corrigé, y_{hij} dans la classe d'imputation k est corrigée de sorte que

$$\tilde{y}_{hij} = \begin{cases} y_{hij} & \text{si la valeur de } y_{hij} \text{ est observée,} \\ E_{A_k}(\tilde{y}_{hij}) - E_{A_k}(y_{hij}) & \text{si la valeur de } y_{hij} \text{ est observée,} \end{cases} \quad (3.1)$$

où E_{A_k} est l'espérance en ce qui concerne la méthode originale d'imputation dans la classe d'imputation k et $E_{A_k}(\tilde{y}_{hij})$ est l'espérance en ce qui concerne la méthode d'imputation corrigée. Cette formule est donnée fondée uniquement sur les données de la $j^{ième}$ répétition de l'enquête. On peut également utiliser la méthode de développement présentée par Rao et Shao (1992) pour le rééchantillonnage par le jackknife et l'imputation par la méthode hot deck pondérée.

L'enregistrement pour l'unité u , nous pouvons ajouter les poids de rééchantillonnage $w_{(u)}^n, r = 1 \text{ à } R$, dans l'exemple où chaque unité d'échantillonnage u . Donc, dans l'exemple où même façon que θ , mais en remplaçant $w_{(u)}^n$ par $w_{(u)}^n$ pour l'unité u ne fait pas partie de la répétition r , alors $w_{(u)}^n = 0$. Une partie ou l'ensemble des poids de rééchantillonnage applicables aux diverses unités incluses dans la répétition seront plus grands que le poids d'échantillonnage, si bien que les unités comprises dans la répétition continueront de représenter l'ensemble de la population. L'utilisation des poids de rééchantillonnage fournis dans le fichier pour calculer les estimations de la variance d'échantillonnage présente les avantages qui suivent :

- Toute statistique, aussi compliquée soit-elle, pouvant être calculée pour l'ensemble de l'échantillon peut aussi facilement l'être pour chaque répétition. La variance d'échantillonnage est alors estimée selon (2.1).
- Les corrections pour tenir compte de la non-réponse totale et de la stratification a posteriori peuvent (et devraient) être réalisées individuellement pour chaque répétition et des unités d'échantillonnage et que l'analyste des données puisse les utiliser sans effort supplémentaire.
- Les corrections apportées aux poids de rééchantillonnage qui figurent dans le fichier peuvent se fonder sur des données auxiliaires auxquelles l'analyste n'a pas nécessairement accès, parfois pour des raisons de confidentialité, ou qu'il pourrait avoir de la difficulté à obtenir ou à utiliser, même si leur consultation n'est pas restreinte.

- Des logiciels d'usage général sont capables de traiter les poids de rééchantillonnage. Deux logiciels qui mettent l'accent sur les méthodes de rééchantillonnage applicables aux données d'enquête sont WesVar de Westat Inc. et VPLX du U.S. Census Bureau. Pour des renseignements sur les logiciels d'analyse des données d'enquête, consulter la page Web [//www.fas.harvard.edu/~stats/survey-soft/survey-soft.html](http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html).

Dans la présente section, nous n'avons pas tenu compte des difficultés que pose l'intégration de la composante de la variance due à l'imputation visant à compenser la non-réponse partielle dans les estimations de la variance. Nous examinons ces difficultés à la section suivante.

Application de l'estimation de la variance selon Rao-Shao en utilisant des poids de rééchantillonnage

MICHAEL P. COHEN¹

RESUME

Pour estimer la variance en tenant compte de l'imputation pour la non-réponse partielle, Rao et Shao (1992) ont lancé une méthode fondée sur un rééchantillonnage corrigé. Shao, Chen et Chen (1998) ont apporté plusieurs perfectionnements à la méthode (particulièrement l'extension du rééchantillonnage par le jackknife de Rao et Shao au rééchantillonnage équilibré une méthode BRR). Dans le présent article, nous examinons comment il est possible d'appliquer ces méthodes en utilisant des poids de rééchantillonnage.

MOTS CLES : Rééchantillonnage équilibré (méthode BRR); rééchantillonnage par le Jackknife; imputation; non-réponse partielle; méthode hot deck pondérée.

1. INTRODUCTION

L'utilisation de poids de rééchantillonnage facilite l'estimation de la variance par les méthodes de rééchantillonnage (Dippo, Fay et Morgenthau 1984). Au cours de la dernière décennie, plusieurs méthodes de ce genre ont été mises au point (Rao et Shao 1992; Shao, Chen et Chen 1998) afin de tenir compte de la variance due à l'imputation pour la non-réponse partielle lors de l'estimation de la variance. Toutefois, la façon d'appliquer ces méthodes de rééchantillonnage corrigé en se servant de poids de rééchantillonnage n'est pas entièrement évidente. Le présent article examine les moyens de le faire. Il se concentre sur la façon de préparer l'ensemble de données pour que les logiciels courants d'estimation de la variance à l'aide de poids de rééchantillonnage puissent être utilisés sans aucune modification. Néanmoins, à l'avant-dernière section, nous formulons certains commentaires quant à l'utilité éventuelle de la modification du logiciel.

2. MÉTHODES DE RÉCHANTILLONNAGE ET POIDS DE RÉCHANTILLONNAGE

Wolter (1985) donne une introduction complète à l'estimation de la variance dans le cas des enquêtes par sondage. Les chapitres 3 et 4 couvrent les deux méthodes par réchantillonnage dont il est question dans le présent article, à savoir le réchantillonnage par le jackknife et le réchantillonnage équilibré (balanced repeated replication). Pour un traitement plus récent et plus poussé de la question, il est conseillé au lecteur de consulter Shao et Tu (1993, chapitre 6). L'estimation de la variance des données d'enquête par réchantillonnage continue de faire l'objet de recherche. Parmi les travaux encore plus récents, mentionnons Brick et Morganstein (1996, 1997), Kott

$$(2.1) \quad \widehat{\langle \theta \rangle}^{\text{var}} = \sum_{R=1}^{\infty} C_{M,R} (\hat{\theta})^{R-2} (\hat{\theta})^2$$

(2001), Rao et Shao (1996, 1999), Rust et Rao (1996), Shao (1996) et Valliant (1996).

La création de sous-ensembles de l'échantillon appelés *répé-
rations*. Le schéma selon lequel sont formées les répétitions
est l'élément qui différencie les deux méthodes. Dans le cas
du rééchantillonnage équilibré, encore appelé méthode du
pondé-échantillon équilibré répété, les répétitions corres-
pondent à environ la moitié des unités de l'échantillon
original; par conséquent, elles sont également appelées
demi-échantillons. Dans le cas du rééchantillonnage par le
jackknife (tel qu'appliqué aux données d'enquête), les
répétitions correspondent habituellement à l'échantillon
original, dont on a supprimé une unité primaire d'échantil-
lonnage (UPB) ou un petit nombre d'UPB appartenant à la
même strate. Dans le cas des deux méthodes, les répétitions
sont considérées comme des échantillons à part entière. Par
conséquent, si θ représente l'estimation d'une certaine
quantité θ fondée sur l'échantillon original, nous pouvons
former une estimation $\theta^{(r)}$ de θ fondée sur la répétition r . Si
nous procédons à R répétitions, nous estimons la variance
d'échantillonnage de θ , var(θ), par

où la constante $C_{M,R}$ dépend uniquement de la méthode de rééchantillonnage M et du nombre de répétitions R . Pour former l'estimation $\hat{\theta}$ de θ , nous utilisons les poids d'échantillonnage. Par exemple, pour estimer un total de population pour une variable particulière y , nous calculons la somme pondérée des valeurs de y . Donc, si y'' et w'' sont les valeurs de y et du poids d'échantillonnage pour l'unité d'échantillonnage n , alors $\hat{\theta} = \sum w'' y''$, où la somme est calculée sur l'ensemble des unités échantillonnées. En plus du poids d'échantillonnage w'' , figurant dans

Nous procédons par induction mathématique. Si $t = 1$,

Par l'hypothèse (2),

$$\text{Cov}_m(Y^t, Y^t) = \alpha_1^2 V^m(Y_0) + \sigma^2 E_m(Y_0)$$

$$= N(\alpha_1^2 v_0 + \sigma^2 \mu_0)$$

$$= V^m(Y^1).$$

Posons maintenant que (9) se vérifie au moment $t - 1$. Soit $E_{t-1} V^m$ et Cov_t l'espérance, la variance et la covariance en conditionnalité par $y_{t-1}^1, R_{t-1}^j, j = 1, \dots, t$. Dans ce cas,

$$E_t(Y^t) = \alpha_1 Y_{t-1}^1$$

et

$$\text{Cov}_t(Y^t, Y^t) = \text{Cov}_t(\alpha_1 Y_{t-1}^1, Y^t)$$

$$= Y_{t-1}^1 \text{Cov}_t(\alpha_1, Y^t)$$

$$= \sigma^2 Y_{t-1}^1$$

où la dernière égalité procède de l'hypothèse (2). Par l'hypothèse d'induction,

$$\text{Cov}_m(Y^t, Y_{t-1}^1) = V^m(Y_{t-1}^1).$$

Alors,

$$\text{Cov}_m(Y^t, Y^t) = \text{Cov}_m[E_t(Y^t), E_t(Y^t)] + E_m[\text{Cov}_t(Y^t, Y^t)]$$

$$= \alpha_1^2 \text{Cov}_m(Y_{t-1}^1, Y_{t-1}^1) + \sigma^2 E_m(Y_{t-1}^1)$$

$$= \alpha_1^2 V^m(Y_{t-1}^1) + \sigma^2 E_m(Y_{t-1}^1)$$

$$= V^m(Y^t).$$

BIBLIOGRAPHIE

BUTANI, S., HARTER, R. et WOLTER, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 523-528.

DIPPO, C.S., FAY, R.E. et MORGANSTEIN, D.H. (1984). Computing variances from complex samples with replicate weights. Dans *Proceedings of the Section on Survey Research Methodology*, American Statistical Association. 489-494.

JUDKINS, D.R. (1990). Fay's method of variance estimation. *Journal of the Official Statistics*, 6, 223-239.

LEE, H., RANCOURT, E. and SÄRDAL C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

RAO, J.N.K. et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et SHAO, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

SHAO, J., et CHEN, Y. (1999). Approximate balanced half samples and related replication methods for imputed survey data. *Sankhya, B, Special Issue on Sample Surveys*, 187-201.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Tableau 5

Résultats de simulation pour la rénumération hebdomadaire																							
Estimation de variance pour le total estimatif												Estimation de la variation											
V_1												V_0											
V_1												V_0											
RM	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM	RM	CV	PC	LM	RB	CV	PC	LM	RB	CV	PC	LM
1	2,0E9	-0,1	9,5E12	-30,7	30,4	81,8	10,3	1,7	41,0	90,0	12,4	17,2	44,3	92,4	13,3	39,8	54,4	94,4	14,6	4,3	48,9	91,0	12,6
2	2,1E9	-0,1	1,7E ¹⁵	-27,2	27,8	84,3	14,1	-3,4	38,7	89,2	16,2	7,9	41,2	91,2	17,1	31,1	48,1	93,5	18,9	3,3	51,5	91,6	16,8
3	2,1E9	-0,1	2,2E ¹⁵	-14,3	34,7	85,6	17,4	1,1	42,2	88,1	18,9	8,0	43,9	89,5	19,5	34,9	51,4	93,5	21,8	2,6	50,4	91,4	19,0
4	2,2E9	-0,1	3,7E ¹⁵	-12,3	40,3	90,1	22,8	6,4	50,6	92,8	25,1	13,8	53,0	94,1	26,0	41,2	63,0	96,1	28,9	-0,9	84,5	92,4	24,2
5	2,2E9	-0,1	5,0E ¹⁵	-16,0	41,6	89,0	25,9	-1,5	51,8	91,4	28,1	5,9	54,8	92,0	29,1	29,3	64,6	94,3	32,2	-5,4	56,0	92,4	27,5
6	2,2E9	-0,1	4,5E ¹⁵	-9,4	44,1	92,0	25,5	-3,8	46,9	92,6	26,3	1,8	48,7	92,8	27,1	27,8	57,8	95,0	30,3	-0,4	54,1	94,2	26,8
7	2,2E9	-0,1	3,5E ¹⁵	-7,3	43,1	92,1	22,8	-0,7	48,3	92,8	23,6	6,8	50,0	93,9	24,5	31,9	57,0	96,4	27,2	-0,0	54,3	95,3	23,7
Estimation de variance pour la variation estimative												Total : total de population											
V_1												V_0											
V_1												V_0											
V_1												V_0											
2	6,4E7	-0,1	1,5E13	-37,6	25,7	85,4	12,2	-8,2	38,4	93,0	14,8	0,2	40,4	94,1	15,5	21,6	47,7	95,8	17,1	5,5	49,2	92,6	15,9
3	3,5E7	-1,6	1,3E13	-31,7	27,9	87,7	11,9	-5,2	42,3	92,2	14,0	2,2	43,8	92,8	14,6	22,3	48,9	94,3	15,9	3,5	43,2	93,5	14,7
4	2,1E7	6,6	2,4E13	-29,5	47,1	86,7	16,5	0,4	63,2	91,9	19,6	6,7	66,2	92,6	20,2	30,7	78,7	95,2	22,4	-4,3	96,9	90,6	19,2
5	2,1E7	-0,4	2,4E13	-40,5	34,1	83,5	15,1	-9,2	55,7	90,5	18,7	-2,4	58,9	92,0	19,4	19,9	69,2	94,9	21,5	3,6	90,0	92,5	19,9
6	1,4E7	2,0	2,3E13	-40,8	31,1	84,4	14,8	-13,5	46,0	91,4	17,8	-6,7	48,9	92,1	18,5	16,8	60,1	94,5	20,7	-4,4	53,0	91,5	18,8
7	1,1E7	-0,1	2,7E13	-40,5	42,0	83,1	16,0	-13,9	56,5	89,2	19,3	-8,7	58,7	90,6	19,9	13,0	68,8	92,8	22,1	-3,7	69,5	90,8	20,4

Total : total de population.
Variation : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimés.
BR : biais relatif = 100 (biases/juste valeur) %.
CV : coefficient de variation = 100 (centeur-type/juste valeur) %.
PC : Probabilité de couvrir de l'intervalle de confiance asymptotique avec variance estimée (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10¹².
* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

6. CONCLUSION ET EXAMEN

Pour les estimateurs avec données d'imputation de la Current Employment Survey (CES), nous proposons un estimateur $v_1 - v_2$ asymptotiquement sans biais et convergent (section 3). S'il est facile de calculer v_1 par la méthode de regroupement en demi-échantillons équilibrés, l'établissement de v_2 comporte des calculs distincts pour les estimateurs non linéaires. Ainsi, nous considérons plusieurs approximations, v_{11} , v_{12} et v_{13} (section 4), et les comparons à $v_1 - v_2$ dans une étude de simulation où un ensemble de données de la CES nous sert de population. Les résultats indiquent que v_{11} et v_{12} ont d'importants biais relatifs imputables au caractère non négligeable du taux d'échantillonnage global (15 %). L'estimateur v_{11} qui est le même que v_1 , mais après intégration d'un taux d'échantillonnage estimé (par le taux de réponse) dans l'application de la méthode de regroupement en demi-échantillons équilibrés, est d'un assez bon rendement. Ainsi, nous recommandons de remplacer $v_1 - v_2$ par v_{11} si le calcul de v_2 est trop complexe. Comme le recours au « taux d'échantillonnage observé » r_{hi}/N_h ne tient pas compte de ce que des données des mois antérieurs soient disponibles sur les non-répondants, il est possible d'améliorer v_{11} en faisant intervenir un taux d'échantillonnage estimé plus fidèle, ce qui peut être, par exemple, la « fraction de données manquantes » de Rubin (1987).

REMERCIEMENTS

Notre étude vise la CES, mais les résultats obtenus sont applicables à toute enquête ayant un plan d'échantillonnage et une méthode d'imputation semblables. Ajoutons qu'une extension au cas où le modèle (2) comprend $y_{1i}^{t-1}, \dots, y_{1i}^{t-s}$ avec un nombre entier $s \geq 2$ est chose simple, bien que le calcul de v_2 (pour un estimateur de variance asymptotiquement valide) soit plus compliqué.

Les auteurs remercient un rédacteur adjoint et deux examinateurs de leurs observations et de leurs suggestions utiles. La recherche de Jun Shao est partiellement financée par la bourse NSF 9803112, 01-02223 de la DMS et de la bourse NSA 904-99-1-0032 de la MDA.

ANNEXE : DÉMONSTRATION DE (4)

Il suffit de démontrer que

$$\text{Cov}^m(X^t, Y^t) = V^m(Y^t).$$

Nous présentons le cas d'une cellule d'imputation unique et d'une égalité $y_{1i}^E = y_{1i}^E$ (emploi). Le cas général peut faire l'objet d'un traitement semblable.

(9)

Tableau 3
Résultats de simulation pour les travailleurs hors personnel de surveillance

Mois	Estimation du total											
	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC
1	5,4E6	-0,1	4,6E7	-33,3	49,7	80,9	7,0	-4,4	66,1	88,1	8,4	8,4
2	5,5E6	-0,1	7,6E7	-30,6	31,4	84,0	9,2	-7,4	41,1	89,4	10,6	10,6
3	5,6E6	-0,1	1,2E8	-23,6	31,2	85,6	12,5	-12,8	41,0	89,5	13,5	13,5
4	5,6E6	-0,1	1,9E8	-19,0	34,5	88,4	15,7	-2,4	43,8	91,7	17,2	17,2
5	5,7E6	-0,1	2,4E8	-18,9	36,8	87,8	17,6	-7,1	45,3	89,7	18,9	18,9
6	5,7E6	0,0	1,8E8	-7,6	41,7	91,8	16,3	-4,7	42,8	92,4	16,6	16,6
7	5,7E6	0,0	1,4E8	-10,9	36,1	91,9	14,1	-7,7	37,2	92,2	14,4	14,4
Estimation de la variance pour la variation estimative												
	V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
Estimation de variance pour la variation estimative												
	V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
1	8,4	-6,0	43,7	92,4	8,8	8,2	48,8	94,4	9,4	9,9	54,8	93,0
2	8,4	-6,0	43,7	92,4	8,8	8,2	48,8	94,4	9,4	9,9	54,8	93,0
3	9,5	-7,2	44,5	91,7	9,8	12,3	49,9	94,1	10,6	10,6	54,8	93,0
4	11,7	-9,9	46,1	90,8	11,7	14,5	51,4	93,4	13,2	13,2	54,8	93,0
5	11,8	-10,0	46,1	90,8	11,7	14,5	51,4	93,4	13,2	13,2	54,8	93,0
6	11,8	-10,0	46,1	90,8	11,7	14,5	51,4	93,4	13,2	13,2	54,8	93,0
7	12,6	-10,8	46,1	90,8	11,7	14,5	51,4	93,4	13,2	13,2	54,8	93,0

Total : variance de la population
Total : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimée.
BR : biais relatif = 100 (biais/juste valeur) %.

CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10⁶.

* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

Tableau 4
Résultats de simulation pour les heures travaillées

Mois	Estimation du total											
	Var*	BR	CV	PC	LM	BR	CV	PC	LM	BR	CV	PC
1	1,9E8	-0,1	5,8E10	-31,5	28,0	79,0	8,0	2,3	44,4	88,3	9,7	12,3
2	2,0E8	-0,1	1,2E11	-30,2	32,8	84,7	11,6	-7,7	40,4	90,6	13,3	13,3
3	2,0E8	-0,1	1,8E11	-23,3	30,0	86,3	14,9	-6,3	36,7	90,3	16,4	16,4
4	2,0E8	0,0	3,2E12	-20,2	35,6	90,2	20,2	-0,5	47,1	93,4	22,6	22,6
5	2,1E8	0,0	4,4E11	-21,2	40,5	88,9	23,6	-7,9	52,3	90,7	25,5	25,5
6	2,1E8	0,0	3,4E11	-10,4	46,3	92,1	22,1	-5,9	48,9	92,2	22,6	22,6
7	2,1E8	0,0	2,3E11	-7,0	40,8	93,0	18,5	-2,2	42,8	93,2	19,0	19,0
Estimation de la variance pour la variation estimative												
	V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
Estimation de variance pour la variation estimative												
	V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
1	11,1	8,0	48,7	90,9	10,0	11,1	53,4	93,6	11,1	11,1	53,4	93,6
2	14,1	9,0	49,1	90,1	14,1	15,2	49,4	94,3	15,2	15,2	49,4	94,3
3	17,1	10,0	49,2	90,7	17,1	18,6	49,6	94,6	18,6	18,6	49,6	94,6
4	22,6	11,1	49,7	91,2	22,6	25,6	49,8	95,3	25,6	25,6	49,8	95,3
5	25,8	12,6	49,9	91,6	25,8	28,8	50,1	95,8	28,8	28,8	50,1	95,8
6	22,9	11,8	49,7	90,9	22,9	25,6	49,7	94,7	25,6	25,6	49,7	94,7
7	18,4	10,0	49,0	90,9	18,4	21,6	49,5	94,8	21,6	21,6	49,5	94,8

Total : variance de la population
Total : différence de population entre le mois en cours et le mois précédent.
Var : variance du total ou de la variation estimée.
BR : biais relatif = 100 (biais/juste valeur) %.

CV : coefficient de variation = 100 (erreur-type/juste valeur) %.
PC : Probabilité de couverture de l'intervalle de confiance asymptotique (en %).
LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10⁶.

* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

Tableau 1
Taille d'échantillon par strate

CTI	Taille de strate	Taux d'échantillon	CTI	Taille de strate	Taux d'échantillon
10, 12-14	567	14	50-51	3631	66
2	433	303	2	3678	183
3	526	1,00000	3	4300	403
4	210	1,00000	4	1831	289
5	165	1,00000	5	833	320
15-17	5055	0,02549	52-59	7084	149
1	129	0,02549	1	7084	0,02103
2	4476	0,12731	2	5701	0,07724
3	5281	0,21854	3	8363	0,12403
4	2111	0,39583	4	4311	0,16915
5	1005	1,00000	5	4087	0,24528
24-25, 32-29	3103	0,02349	60-62, 67	1384	17
2	3905	0,08475	2	971	0,03906
3	891	0,13966	3	1529	0,07500
4	4273	0,24242	4	981	0,06818
5	2127	0,51351	5	728	0,10000
20-23, 26-31	1754	0,02276	63-64	1364	0,01119
1	40	0,02276	1	652	0,03125
2	1953	0,06564	2	754	0,11538
3	3591	0,14599	3	87	0,11538
4	3108	0,19167	4	435	0,11110
5	1041	0,30189	5	344	0,16667
40-49	1648	0,01902	7, 70-99	9641	230
1	1648	0,01902	1	9641	0,02385
2	1463	0,06918	2	6701	0,09602
3	1988	0,11111	3	7833	0,16275
4	1171	0,18033	4	4839	0,27215
5	759	0,14286	5	4352	0,47500

Tableau 2
Résultats de simulation pour la variable de l'emploi

Mois	Total*	Estimation du total										Estimation de variance pour le total estimatif																
		BR	Var*	BR	CV	PC	LM	V_{t0}	BR	CV	PC	LM	V_{t1}	BR	CV	PC	LM	V_{t1}	BR	CV	PC	LM	$V_{t1} - V_{t2}$					
1	6,7E6	0,0	5,5E7	-37,0	47,6	85,3	7,7	-4,1	67,5	92,3	9,2	4,9	69,8	93,1	9,6	19,5	76,1	95,1	10,3	7,4	67,4	92,8	9,7	4,4	49,1	92,3	12,1	$V_{t1} - V_{t2}$
2	6,8E6	0,0	8,8E7	-34,3	28,8	86,9	9,6	-7,3	40,4	92,6	11,4	0,9	42,9	93,6	12,4	15,3	47,6	94,7	12,7	4,4	49,1	92,3	12,1	4,4	49,1	92,3	12,1	$V_{t1} - V_{t2}$
3	6,9E6	0,0	1,4E8	-26,1	30,4	88,2	12,9	-4,1	42,3	91,8	14,7	1,4	44,2	92,9	15,1	18,8	49,9	94,8	16,3	3,6	50,5	90,8	15,2	3,6	50,5	90,8	15,2	$V_{t1} - V_{t2}$
4	6,9E6	0,0	2,1E8	-22,5	32,9	89,3	16,1	-2,4	44,0	92,1	18,1	3,8	46,3	92,7	18,7	22,3	53,1	94,7	20,3	2,7	51,3	91,4	18,6	-4,7	54,2	90,9	20,3	$V_{t1} - V_{t2}$
5	6,9E6	0,0	2,7E8	-21,9	35,0	88,3	18,4	-7,7	45,2	90,9	20,0	-1,1	47,9	92,0	20,7	16,2	55,6	94,4	22,4	-4,7	54,2	90,9	20,3	-4,7	54,2	90,9	20,3	$V_{t1} - V_{t2}$
6	6,9E6	0,0	2,0E8	-8,8	40,5	91,7	17,1	-5,2	41,7	91,9	17,4	0,0	43,6	93,1	17,9	19,7	51,8	95,5	19,6	-3,1	52,5	90,5	17,6	-6,6	42,4	92,7	15,0	$V_{t1} - V_{t2}$
7	6,9E6	0,0	1,5E8	-12,4	34,8	91,8	14,5	-8,6	36,1	92,5	14,8	-2,0	38,3	93,6	15,3	16,8	45,0	96,2	16,7	-6,6	42,4	92,7	15,0	-6,6	42,4	92,7	15,0	$V_{t1} - V_{t2}$
2	8,0E4	-0,1	6,1E7	-43,0	25,4	84,9	7,5	-11,3	41,4	92,3	9,3	-4,5	43,9	93,7	9,7	9,4	48,7	95,6	10,3	8,6	51,7	93,5	10,3	8,6	51,7	93,5	10,3	$V_{t1} - V_{t2}$
3	9,7E4	-1,8	7,4E7	-35,0	31,7	85,0	8,7	-8,5	46,0	90,5	10,4	-3,2	47,7	91,0	10,7	11,7	53,1	93,4	11,5	-3,1	48,8	90,9	10,7	-3,1	48,8	90,9	10,7	$V_{t1} - V_{t2}$
4	1,8E4	2,9	1,1E8	-31,8	42,3	87,4	11,0	-0,9	60,6	93,1	13,2	4,9	63,2	93,6	13,6	25,0	73,5	95,9	14,8	-2,5	47,7	89,9	13,1	-2,5	47,7	89,9	13,1	$V_{t1} - V_{t2}$
5	4,4E4	3,4	1,1E8	-41,9	34,5	83,1	10,1	-10,8	57,3	91,4	12,5	-4,9	60,4	92,3	12,9	13,2	69,4	94,6	14,1	0,8	94,1	93,1	13,3	0,8	94,1	93,1	13,3	$V_{t1} - V_{t2}$
6	-1,1E4	9,3	1,1E8	-41,0	29,9	84,1	10,2	-12,6	42,0	91,1	12,4	-6,4	44,2	93,0	12,8	9,4	50,2	94,6	13,9	-4,1	53,9	93,0	13,0	-4,1	53,9	93,0	13,0	$V_{t1} - V_{t2}$
7	1,6E3	3,2	1,2E8	-43,8	38,4	82,9	10,4	-15,9	57,5	89,6	12,7	-11,3	60,1	90,5	13,1	5,6	69,9	92,6	14,2	-0,2	75,5	90,0	13,8	-0,2	75,5	90,0	13,8	$V_{t1} - V_{t2}$

Var : variance du total ou de la variation estimés.

BR : biais relatif = 100 (biais/juste valeur) %.

CV : coefficient de variation = 100 (erreur-typique/juste valeur) %.

PC : Probabilité de couverture de l'intervalle de confiance asymptotique avec variance estimée (en %).

LM : (Largeur moyenne de l'intervalle de confiance asymptotique) / 10³.

* : Notation scientifique (Exemple 6 700 000 est 6,7E6).

4. L'estimateur de variance v_{11} , qui est le même que v_{11}^p mais après intégration des taux d'échantillonnage n_h/N_h (section 4), présente en général un biais relatif en valeur négative. Ce biais négatif peut être important, plus particulièrement dans l'estimation de variance pour les variations mensuelles.
5. L'estimateur de variance v_{11} , qui est le même que v_{11}^p mais après remplacement des taux d'échantillonnage n_h/N_h par $r_{h,1}/N_h$, est d'un bon rendement dans l'étude de simulation, bien que n'étant pas asymptotiquement sans biais (section 4). Son biais relatif est important dans quelques cas, entre autres dans les estimations de variance pour le total de la rémunération hebdomadaire dans les mois 1 et 4, pour le total des heures dans le mois 1 et pour la variation d'emploi dans le mois 7. Dans bien des cas toutefois, v_{11} est encore d'un meilleur rendement que l'estimateur asymptotiquement sans biais $v_{11} - v_{12}$.
- Voici en résumé les résultats de simulation des intervalles de confiance pour ce qui est de la PC et de la LM.

1. La PC de l'intervalle de confiance pour l'estimateur simple de variance v_{10} se situe bien en deçà dans la plupart des cas du niveau nominal de 95 %.
2. La PC de l'intervalle de confiance pour l'estimateur de variance asymptotiquement valide $v_{11} - v_{12}$ est comprise entre 90 % et 93 % dans la plupart des cas. Tel est souvent le cas avec un estimateur de variance asymptotiquement valide : le biais relatif est petit, mais la PC de l'intervalle de confiance correspondant est inférieure au niveau nominal. La raison en est peut-être que la convergence de distribution (normalité asymptotique qui est à la base même des intervalles de confiance asymptotiques) exige plus en taille d'échantillon que la convergence du deuxième moment (estimation de variance).

3. En ce qui concerne la PC, l'intervalle de confiance pour v_{11} est le meilleur, peut-être parce que la surestimation de variance compense la sous-couverture de l'estimation d'intervalle. La largeur moyenne de l'intervalle pour v_{11} peut être bien supérieure à celle d'autres intervalles, notamment pour la variable de la rémunération hebdomadaire.
4. La PC de l'intervalle de confiance pour v_{11} , qui n'est pas asymptotiquement valide, est semblable à celle de l'intervalle pour $v_{11} - v_{12}$.

population P . Chaque unité $i \in P$ a un vecteur $y_i = (y_{i1}, y_{i2}, \dots, y_{it}, \dots, y_{i7}, y_{i11}, y_{i12}, y_{i13}, y_{i14}, y_{i15})$, bien que toutes les valeurs de y_i soient disponibles (dans les dossiers administratifs). Dans la simulation, nous obtenons l'échantillon 5 par échantillonnage aléatoire simple stratifié $\{y_i\}$ pour une taille de 23 092 unités de P selon la répartition indiquée au tableau 1. Dans cette simulation, les indicateurs de réponse de $\{y_i\}$ sont issus d'un autre échantillonnage aléatoire simple stratifié (indépendant) $\{r_i\}$ d'unités de P . Ainsi, les non-répondants sont des éléments aléatoires qui se répartissent selon les valeurs des r_i dans l'ensemble de données P , mais en toute indépendance des y_i .

Après obtention des non-répondants et des données d'échantillon, il y a eu imputation à l'égard des premiers comme le décrit la section 2. Nous avons calculé les totaux mensuels y'_t et les variations mensuelles $y'_t - y'_{t-1}$ estimés en fonction des données d'imputation. Le calcul de leurs estimateurs de variance $v'_{11}, v'_{12}, v'_{13}, v'_{14}, v'_{15}$ et $v'_{11} - v'_{12}$ décrit les sections 3 et 4. À des fins de comparaison, nous avons enfin calculé le simple estimateur de variance v_{10} par assimilation des valeurs d'imputation aux données d'observation.

Les tableaux 2 à 5 présentent respectivement pour 4 variables les valeurs obtenues en 1 000 simulations des biais relatifs (BR) et des variances (Var) des totaux y'_t et des variations $y'_t - y'_{t-1}$ estimés, du BR et du coefficient de variation (CV) des estimateurs de variance pour y'_t et $y'_t - y'_{t-1}$, de la probabilité de « couverture » (PC) des intervalles de confiance approximatifs à 95 % (sous la forme :

estimation $\pm 1,96 \sqrt{\text{variance estimative}}$)

et de la largeur moyenne (LM) des intervalles. Les erreurs-types estimées de simulation sont de 2 % pour le BR, le CV et la LM et de 0,5 % pour la PC.

Aux tableaux 2 à 5, les biais relatifs des estimateurs des totaux et des variations mensuels sont négligeables pour toutes les variables. Voici en résumé les résultats de simulation pour les estimateurs de variance (BR et CV).

1. Comme on pouvait le prévoir, l'estimateur simple de variance v_{10} est entaché d'un important biais relatif en valeur négative.
2. L'estimateur de variance asymptotiquement non biaisé $v_{11} - v_{12}$ est d'un bon rendement général. Son biais relatif est toujours de moins de 10 % en valeur absolue. Souvent, il n'atteint pas les 5 %.
3. L'estimateur de variance v_{11} est entaché dans tous les cas d'un grand biais relatif en valeur positive, ce qui indique que le terme v_{12} n'est pas négligeable dans la CES où le taux d'échantillonnage global n/N est d'environ 15 %.

par un estimateur non linéaire comme \hat{Y}'_H/\hat{Y}'_H (répondants y . Le juste estimateur de variance pour \hat{Y} est alors $v_1 - v_2$ avec

$$v_1 = \frac{N^2 \hat{U}^2 s_d^2}{n} + \frac{r}{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2}$$

et

$$v_2 = N \hat{U} s_d^2 + 2N \hat{U} \hat{R} s_{dx} + N \hat{R}^2 s_x^2,$$

où $s_d^2 = (r - 1)^{-1} \sum a_i (y_i - \bar{R} x_i)^2$, $s_{dx} = (r - 1)^{-1} \sum a_i x_i (y_i - \bar{R} x_i)$, et s_x^2 est la variance d'échantillon fondée sur les x_i . De plus,

$$\hat{v}_1 = \left(1 - \frac{r}{N}\right) \left(\frac{N^2 \hat{U}^2 s_d^2}{n} + \frac{r}{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2} \right)$$

$$= v_1 - \frac{r}{nN \hat{U}^2 s_d^2} - 2N \hat{U} \hat{R} s_{dx} - N \hat{R}^2 s_x^2$$

et

$$\hat{v}_1 = \left(1 - \frac{r}{N}\right) \left(\frac{N^2 \hat{U}^2 s_d^2}{n} + \frac{r}{2N^2 \hat{U} \hat{R} s_{dx} + N^2 \hat{R}^2 s_x^2} \right)$$

$$= v_1 - N \hat{U}^2 s_d^2 - \frac{r}{2rN \hat{U} \hat{R} s_{dx} + rN \hat{R}^2 s_x^2}.$$

Comme $v_1 - v_2$ est asymptotiquement sans biais, le biais de $v_1 = v_1$ est du même ordre que v_2 et toujours non négatif. Le biais de $\hat{v}_1 = \hat{v}_1$ est du même ordre que

$$N \hat{U} s_d^2 \left(1 - \frac{r}{nN}\right) = -N \hat{U} s_d^2 \frac{r}{(1 - a_i) x_i} \sum a_i x_i$$

et toujours non positif. Le biais de $\hat{v}_1 = \hat{v}$ est du même ordre que

$$N \hat{U} (1 - \hat{U}) s_d^2 + \left(1 - \frac{r}{n}\right) \left(2N \hat{U} \hat{R} s_{dx} + N \hat{R}^2 s_x^2\right). \quad (8)$$

Le biais en (8) est non négatif si $s_{dx} \geq 0$ et $\hat{U} \approx 1$ (ce qui se vérifie lorsque a_i est indépendant de x_i).

5. CERTAINS RÉSULTATS DE SIMULATION

Pour étudier plus avant les biais des estimateurs de variance v_1 , \hat{v}_1 et \hat{v} , nous avons effectué une étude de simulation à l'aide d'un ensemble de données de la CES (des années 1980) comprenant 149 044 unités comme

$$w_{(r)}' = \begin{cases} (1 + 0,5 \sqrt{1 - n_h/N_h}) w_i & \text{si l'unité est dans le } r^{\text{ème}} \text{ demi-échantillon} \\ (1 - 0,5 \sqrt{1 - n_h/N_h}) w_i & \text{si l'unité n'est pas dans le } r^{\text{ème}} \text{ demi-échantillon,} \end{cases} \quad (7)$$

remplacé par l'observation (voir les résultats de simulation à la section 5). Si v_1 surestime la juste variance $V(\hat{Y}'_t - \hat{Y}_t)$ et que \hat{v}_1 la sous-estime, un moyen terme est possible où on remplace le taux d'échantillonnage n_h/N_h dans (7) par le taux d'échantillonnage estimé $r_{h,i}/N_h$, où $r_{h,i}$ est le nombre de répondants dans la strate h pour le mois t . Soit \hat{v}_1 l'estimateur de variance obtenu par (5) et (7), mais avec remplacement de n_h/N_h par $r_{h,i}/N_h$. Dans ce cas,

$$\hat{v}_1 \leq \hat{v}'_1 \leq v_1.$$

Les trois estimateurs de variance sont asymptotiquement sans biais et approximativement égaux si n/N est négligable. Pour un n/N non négligeable toutefois, ils sont asymptotiquement enchaînés d'un biais. Pour dégrader l'ordre de grandeur des biais de \hat{v}_1 , \hat{v}'_1 et v_1 , nous considérons le cas le plus simple où il n'y a pas de strates et où $t = 1$. Soit $y_i = y_{0,i}$, $x_i = y_{0,i}$ et

$$\hat{Y} = \sum a_i y_i + \sum (1 - a_i) \hat{R} x_i,$$

où $a_i = 1$ si y_i est un répondant et $a_i = 0$ dans les autres cas. $\hat{R} = \sum a_i y_i / (\sum a_i x_i)$ et toutes les sommes sont sur $i \in S$. Soit $\hat{U} = (\sum x_i / n) / (\sum a_i x_i / r)$, où r est le nombre de

(comme nous l'avons décrit à la section précédente) sont asymptotiquement sans biais à l'égard de l'espérance conjointe du modèle (2) et de l'échantillonnage dans la population finie.

Dans la CES, les cellules d'imputation sont des unions de strates et, ainsi,

$$\sum_{i \in S \cap P_k} w_i = M_k, \quad k = 1, \dots, K,$$

où M_k est le nombre d'unités de population dans la $k^{\text{ème}}$ cellule d'imputation P_k . Les X_i sont donc conditionnellement sans biais à l'égard de l'espérance du modèle (étant donné S), c'est-à-dire :

$$E_m(X_i - Y_i) = 0.$$

3. ESTIMATION DE VARIANCE

Soit E_s et V_s l'espérance et la variance respectives d'échantillonnage et V_i la variance globale. Ainsi,

$$V(X_i - Y_i) = E_s[V_m(X_i - Y_i)] + V_s[E_m(X_i - Y_i)] \quad (3)$$

puisque $E_m(X_i - Y_i) = 0$. Nous démontrons en outre en annexe que

$$V_m(X_i - Y_i) = V_m(X_i) - V_m(Y_i). \quad (4)$$

À noter que (4) est évident en cas de non-réponse.

À cause de (3), l'estimation de $V(X_i - Y_i)$ est la même que celle de $V_m(X_i - Y_i)$. De plus, nous pouvons, à cause de (4), calculer d'abord les estimateurs v_{i1} et v_{i2} de $V_m(X_i)$ et $V_m(Y_i)$ respectivement, puis prendre la différence et $V_m(X_i) - V_m(Y_i)$ comme notre estimateur de variance pour $X_i - Y_i$. Comme $V_m(X_i)$ est une variance conditionnelle étant donné S , nous n'avons pas à tenir compte des taux d'échantillonnage n_h/N_h dans l'estimation de $V_m(X_i)$. Si nous considérons d'abord l'estimation de $V_m(X_i)$. Si nous pouvons calculer une formule approximative de $V_m(X_i)$, nous pourrions directement estimer cette valeur par substitution. Il reste que la forme explicite de X_i est fort complexe lorsque i n'est pas petit, d'où la grande difficulté de calculer $V_m(X_i)$. Ainsi, dans la CES, nous adoptons une méthode de regroupement en demi-échantillons où la correction ou (réimputation) de Rao et Shao (1992) permet de tenir compte de l'imputation. Il s'agit plus précisément d'échantillons dans chaque strate. On crée des demi-échantillons R à l'aide d'une matrice d'Hadamard;

$H + 1 \leq R \leq H + 4$ est le nombre de strates. Pour le

$$v_{i1} = \frac{4}{R} \sum_{r=1}^R \left(Y_{(i)}^r - \frac{1}{R} \sum_{r=1}^R Y_{(i)}^r \right)^2. \quad (5)$$

Il convient de noter que, si nous prenons 0,5 au lieu de 1 dans l'élaboration de la pondération $w_i^{(r)}$, c'est en application de la méthode de Fay (Dippo, Fay et Morgansstein 1984; Judkins 1990; Rao et Shao 1999). Asymptotiquement, v_{i1} est sans biais et converge pour $V_m(X_i)$ (Shao, Chen et Chen 1998; Rao et Shao 1999; Shao et Chen 1999).

Considérons maintenant l'estimation de $V_m(Y_i)$. Dans le modèle (2),

$$V_m(Y_i) = \sum_{k=1}^K M_k v_{ik},$$

ce qui est de l'ordre $O(N)$, où N est la taille de la population P . D'ordinaire, $V_m(Y_i)$ est de l'ordre $O(N^2/n)$, où $n = \sum_h n_h$ est la taille d'échantillon. $V_m(X_i)/V_m(Y_i)$ est donc de l'ordre $O(n/N)$, et il est inutile d'estimer $V_m(X_i)$ si n/N est négligeable (bien que certains taux d'échantillonnage n_h/N_h ne le soient pas).

Dans la CES cependant, n/N est d'environ 15 %, valeur non négligeable. Ainsi, l'estimation de $V_m(X_i)$ est nécessaire. Un estimateur asymptotiquement sans biais et convergent de $V_m(X_i)$ sera

$$v_{i2} = \sum_{k=1}^K M_k s_{k2}^2. \quad (6)$$

où s_{k2}^2 est la variance habituelle d'échantillonnage selon les répondants y_{ik} dans la $k^{\text{ème}}$ cellule d'imputation.

4. ESTIMATEURS APPROXIMATIFS DE VARIANCE

On peut voir à la section 3 qu'un juste estimateur de variance de X_i est $v_{i1} - v_{i2}$, où v_{i1} et v_{i2} sont respectivement donnés par (5) et (6). Bien qu'on puisse facilement étendre la détermination de v_{i1} au cas où X_i est remplacé

interstrates. Les taux d'échantillonnage n_h/N_h ne sont pas nécessairement négligeables. Pour un certain nombre de strates se caractérisant par une forte taille des établissements, $n_h = N_h$. Soit S l'échantillon en question. Pour $t \in S$ et le mois $t = 0, 1, \dots, T$, nous observons (s'il n'y a pas non-réponse) le nombre de travailleurs ($y_{t,i}^E$), le nombre de travailleurs hors personnel de surveillance ($y_{t,i}^W$), le nombre de heures travaillées ($y_{t,i}^H$) et la rémunération hebdomadaire ($y_{t,i}^P$). Soit $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$ ou $y_{t,i}^P$. Dans la CES, les grands paramètres d'intérêt sont les totaux de population $X_t = \sum_{i \in P} y_{t,i}$, $t = 1, \dots, T$. Comme on peut tirer chaque année ces totaux des dossiers administratifs, nous supposons sans perte de généralité que X_0 est connu. S'il n'y a pas non-réponse, nous estimons X_t par un estimateur de rapport

$$X_t = X_0 \sum_{i \in S} w_i y_{t,i} / \sum_{i \in S} w_i y_{0,i}, \quad t = 1, \dots, T, \quad (1)$$

où w_i est la valeur de pondération d'enquête de la $i^{\text{ème}}$ unité échantillonnée dans la $h^{\text{ème}}$ strate.

Dans notre recherche, nous appliquons, à partir du mois 1, la méthode d'imputation proposée par Butani, Harter et Wolter (1997) (voir la description qui suit) à l'égard des non-répondants. L'imputation se fait à l'intérieur d'une cellule d'imputation, c'est-à-dire dans une strate ou une union de strates. Les valeurs imputées des mois 1, ..., $t-1$ sont reprises en imputation pour les non-répondants du mois t , sauf si les non-répondants des mois 1, ..., $t-1$ deviennent répondants avant le mois t .

1. Nombre de travailleurs. Si $y_{t,i}^E$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^E = \hat{a}_t \tilde{y}_{t-1,i}^E$$
 où $\tilde{y}_{t-1,i}^E = y_{t-1,i}^E$ (valeur de déclaration) si $y_{t-1,i}^E$ est disponible au mois t , $\tilde{y}_{t-1,i}^E$ étant alors une valeur d'imputation,

$$\hat{a}_t = \frac{\sum_{j \in R_t} w_j y_{t,j}^E}{\sum_{j \in R_t} w_j y_{t-1,j}^E},$$

et où R_t est l'ensemble des unités déclarantes des mois t et $t-1$.

2. Nombre de travailleurs hors personnel de surveillance. Si $y_{t,i}^W$ est non-répondant, il y a imputation par

$$\tilde{y}_{t,i}^W = y_{t-1,i}^W \tilde{y}_{t-1,i}^E / y_{t-1,i}^E,$$
 où $\tilde{y}_{t-1,i}^W$ est défini comme $\tilde{y}_{t-1,i}^E$

3. Nombre d'heures travaillées. Si $y_{t,i}^H$ est non-répondant, il y a imputation par

$$(a_{t,i}^H, \dots, a_{t,i}^H).$$

Après imputation à l'égard des non-répondants, on établit les totaux mensuels estimatifs par la formule (1) en assimilant les valeurs d'imputation à des données de déclaration.

Posons que la population P se divise en K cellules d'imputation disjointes P^1, \dots, P^K et que, pour chaque k ,

$$y_{t,i} = \alpha_{t,k} y_{t-1,i} + \sqrt{y_{t-1,i}} e_{t,i},$$

$$E_m(y_{t,i}) = \mu_{t,k}, \quad E_m(e_{t,i}) = 0, \quad i \in P^k, t = 1, 2, \dots,$$

$$V_m(y_{t,i}) = v_{t,k}, \quad V_m(e_{t,i}) = \sigma_{t,k}^2, \quad (2)$$

où $y_{t,i}$ désigne tout $y_{t,i}^E, y_{t,i}^W, y_{t,i}^H$ ou $y_{t,i}^P$, E_m et V_m sont respectivement l'espérance et la variance (en valeur marginale) du modèle, $\alpha_{t,k}$ et $\sigma_{t,k}^2$ sont des paramètres inconnus, les $e_{t,i}$ sont indépendants et identiquement distribués et les deux opérations $\{y_{t,i}\}$ et $\{e_{t,i}\}$ sont indépendantes. Nous posons aussi que, dans chaque P^k , l'indicateur de réponse $a_{t,i}$ (= 1 si $y_{t,i}$ est répondant et = 0 dans les autres cas) et $y_{t,i}$ sont indépendants, étant donné le mécanisme de réponse sans confusion (Lee, Ranacourt et Särndal 1994), $a_{t,i}$ et $y_{t,i}$ sont dépendants, mais par $y_{t-s,i}, a_{t-s,i}, s = 1, 2, \dots, t$. C'est un cas plus général que celui de l'hypothèse d'indépendance de $(y_{1,i}, \dots, y_{t,i})$ et de $(a_{1,i}, \dots, a_{t,i})$. Posons enfin que les indicateurs de réponse d'unités différentes sont indépendants. Dans ces hypothèses, les estimateurs X_t par les données d'imputation

Estimation de variance dans le cadre de la « Current Employment Survey »

JUN SHAO et SHAIL BUTANI¹

RÉSUMÉ

Comme dans la plupart des autres enquêtes, il y a souvent un phénomène de non-réponse dans la « Current Employment Survey » effectuée tous les mois aux États-Unis par le Bureau of Labor Statistics (BLS). Dans un mois quelconque, les estimateurs seront plus efficaces si on procède à une imputation à l'aide des données déclarées des mois antérieurs au lieu de ne pas tenir compte des répondants et de modifier la pondération de l'enquête en conséquence. Il faut cependant dire que l'imputation influe aussi sur l'estimation de variance. Si on traite les valeurs imputées comme des valeurs déclarées et qu'on applique une méthode de type d'estimation de variance, les estimateurs de variance seront entachés d'un biais négatif. Dans le présent article, nous proposons un certain nombre de ces estimateurs par regroupement en demi-échantillons équilibrés et par réimputation afin de tenir compte de l'imputation. Nous présentons certains résultats de simulation pour le rendement d'échantillonnel pour population finie des estimateurs par données d'imputation et leurs estimateurs de variance.

MOTS CLÉS : Demi-échantillons équilibrés; taux d'échantillonnage non négligeables; imputation de rapport; échantillonnage stratifié.

1. INTRODUCTION

La « Current Employment Survey » (CES), communément appelée enquête sur la rémunération, est menée tous les mois par le Bureau of Labor Statistics (BLS) aux États-Unis. De mois en mois, on recueille par ce moyen des données auprès des établissements, et ce, par divers systèmes automatisés : interviews téléphoniques assistées par ordinateur, entrée de données sur clavier téléphonique, télécopie, transmission électronique de données, poste, etc. Les principales variables de l'enquête sont l'emploi, les travailleurs de la production (sans le personnel de surveillance), leurs heures de travail et leurs gains dans les établissements on agricoles. Chaque année, on tire des chiffres de dénombrement des travailleurs des dossiers administratifs du régime d'assurance-chômage.

La non-réponse est un phénomène fréquent dans la CES. Dans un mois quelconque, on obtient des estimateurs d'enquête plus efficaces par une imputation à l'aide des données déclarées des mois antérieurs que par une prise en compte des seules données de déclaration du mois en cours et une modification de la pondération de l'enquête. Cette constatation vaut particulièrement pour la CES, puisque le taux de non-réponse y est de 60 % à 80 % et qu'environ 60 % des non-répondants d'un mois peuvent devenir disponibles un ou plusieurs mois après, d'où la possibilité d'utiliser ces données dans un mois à venir comme « données déclarées des mois antérieurs » (données chronologiques).

On sait bien toutefois que, si on traite des valeurs d'imputation comme des valeurs de déclaration et applique une méthode de type d'estimation de variance, les estimateurs

de variance seront entachés d'un biais (souvent négatif). Il est possible d'établir des estimateurs de variance valides par certaines hypothèses posées au sujet des plans d'échantillonnage, des méthodes d'imputation et des mécanismes de réponse (et parfois aussi des modèles qui produisent les données).

Dans le présent article, nous étudions l'estimation de variance aux fins de la CES. Après avoir décrit le plan d'échantillonnage et la méthode d'imputation de cette enquête à la section 2, nous dégageons à la section 3 des estimateurs de variance valides (c'est-à-dire asymptotiquement non biaisés et convergents) pour les estimateurs d'enquête par données imputées. Pour simplifier le calcul des estimateurs de variance, nous proposons des approximations à la section 4 et en étudions les propriétés par simulation à la section 5. Nous tirons un certain nombre de conclusions à la section 6. Bien que notre étude ait la CES pour objet, ses résultats en sont, à notre avis, généralisables à toute enquête ayant un plan d'échantillonnage et une méthode d'imputation semblables.

2. PLAN D'ÉCHANTILLONNAGE ET IMPUTATION

La CES applique le plan suivant d'échantillonnage probabiliste stratifié. Soit P une population finie constituée d'un ensemble d'établissements $\{1, \dots, N\}$ et stratifiée selon l'industrie et la taille des établissements. Dans la strate h ^{ème}, nous prélevons un échantillon de taille $n_h \geq 2$ sans remise sur N_h unités de population par application indépendante d'un plan d'échantillonnage probabiliste

¹ Jun Shao, Département de statistique, Université du Wisconsin, Madison WI 53706; Shail Butani, Statistical Methods Division, Bureau of Labor Statistics, Washington, D.C. 20212.

12. CONCLUSIONS

L'estimateur que nous proposons est un des rares estimateurs qui soit à la fois sans biais, linéaire, qui utilise de l'information auxiliaire, et qui est calé sur la taille de la population. Ce nouvel estimateur est robuste par rapport à la fenêtre q . Il peut prendre en compte une information auxiliaire ayant une relation non-linéaire avec l'estimateur par la régression. Il peut prendre en compte une information auxiliaire ayant une relation non-linéaire avec la variable d'intérêt. La plupart des simulations semblent montrer que la largeur de la fenêtre n'a pas beaucoup d'impacts sur la précision d'un estimateur de moyenne. Cependant, il apparaît aussi qu'une petite largeur de fenêtre n'est pas pénalisante, même si il n'y a pas de dépendance entre la variable auxiliaire et la variable d'intérêt. Plus q est petit, plus le calage sera serré, et l'estimateur de la variance perdu dans chaque post-strate. Le choix de q doit donc tenir compte de ce problème.

Il existe beaucoup d'autres méthodes permettant d'utiliser l'information donnée par une fonction de répartition (voir Ren 2000) pour améliorer un estimateur. Les résultats que nous avons présentés se réduisent aux plans simples, mais nous pensons qu'ils sont importants, au même titre que la post-stratification est importante comme cas particulier des techniques de calage. En effet, la post-stratification est un des rares exemples où l'on peut montrer avec exactitude que le calage correspond à une approche conditionnelle. De plus, notre approche peut être vue comme un calage sur une fonction de répartition fournissant un estimateur sans biais. Une bonne technique générale de calage sur la répartition devrait donc retrouver dans les plans simples la méthode que nous avons présentée.

BIBLIOGRAPHIE

- Favre, deux arbitres et un éditeur associé pour leurs commentaires constructifs qui ont permis d'améliorer considérablement cet article.
- Nous remercions Jean-Claude Deville et Anne-Catherine Deville, j.-C. (1995). *Estimation de la variance du coefficient de Gini mesuré par sondage*. INSEE Méthode, document de travail, Méthodologie F9510.
- DEVILLE, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête*, 25, 219-230.
- DEVILLE, J.-C., et SÄRNÄDAL, C.-E. (1992). Calibration estimateurs in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ESTEVAO, V., HIDIROGLOU, M.A. et SÄRNÄDAL, C.-E. (1995). Methodological principle for a generalized estimation system in Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- REN, R. (2000). *Estimation par calage sur la répartition*. Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.
- TILLÉ, Y. (1998). Estimation in surveys using conditional inclusion probabilities : simple random sampling. *International Statistical Review*, 66, 303-322.
- TILLÉ, Y. (1999a). Sur la détermination a posteriori des bornes des post-strates. Dans *Les sondages*, (Eds. G. Brossier, et A.-M. Dussaix), Dunod, 202-208.
- TILLÉ, Y. (1999b). Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles : enquêtes à plan d'échantillonnage complexe. *Techniques d'enquêtes*, 25, 57-66.
- WU, C., et SITTER, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289-307.

REMERCIEMENTS

Ces deux estimateurs de variances sont à évaluer par des simulations.

II. SIMULATIONS POUR LES ESTIMATEURS DE VARIANCE

Les simulations présentées dans le tableau (5) sont basées sur des populations de taille $N = 100$, qui sont générées au moyen de variables aléatoires indépendantes normales. Pour chaque cas étudié, on donne la valeur de q , et le coefficient de corrélation entre la variable d'intérêt Y_k et le rang R_k de la variable auxiliaire X_k . La borne b est définie en prenant la partie entière de $q/2+1$. L'objectif étant de valider l'estimateur de variance, on tire 3 000 échantillons de taille $n = 20$ pour chaque simulation, et on compare la variance estimée par les simulations de l'estimateur calé $V_{st}(Y_c)$ aux espérances sous les simulations des deux estimateurs de variance notées $E_{st}(V_a(Y_c))$, $\alpha = 1, 2$. Les deux dernières colonnes des tableaux présentent le biais relatif défini par

$$BR_{st} V_a(Y_c) = \frac{E_{st} V_a(Y_c) - V_{st}(Y_c)}{V_{st}(Y_c)}, \alpha = 1, 2.$$

Les simulations montrent que les deux estimateurs proposés surestiment la variance. La surestimation semble diminuer quand q croît. L'estimateur $V_2(Y_c)$ a manifestement le plus petit biais. On préconisera donc l'utilisation de

Tableau 5

Résultats des simulations

Corrélation: 0,802									
q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$	$BR_{st} V_1(Y_c)$	$BR_{st} V_2(Y_c)$	q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$
4	0,045	0,065	0,054	0,444	0,200	4	0,048	0,066	0,059
5	0,045	0,066	0,070	0,467	0,267	5	0,044	0,057	0,054
6	0,056	0,076	0,070	0,357	0,250	6	0,044	0,056	0,051
7	0,058	0,079	0,059	0,362	0,107	7	0,044	0,054	0,051
8	0,063	0,088	0,087	0,397	0,381	8	0,045	0,052	0,048
Corrélation: 0,481									
q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$	$BR_{st} V_1(Y_c)$	$BR_{st} V_2(Y_c)$	q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$
4	0,281	0,471	0,363	0,676	0,292	4	0,297	0,420	0,356
5	0,281	0,471	0,363	0,414	0,199	5	0,279	0,363	0,316
6	0,267	0,342	0,281	0,281	0,133	6	0,267	0,342	0,281
7	0,282	0,327	0,281	0,160	-0,004	7	0,282	0,327	0,281
Corrélation: 0,111									
q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$	$BR_{st} V_1(Y_c)$	$BR_{st} V_2(Y_c)$	q	$V_{st}(Y_c)$	$E_{st} V_1(Y_c)$	$E_{st} V_2(Y_c)$
4	0,281	0,471	0,363	0,676	0,292	4	0,281	0,471	0,363
5	0,281	0,471	0,363	0,414	0,199	5	0,281	0,471	0,363
6	0,279	0,363	0,316	0,301	0,133	6	0,279	0,363	0,316
7	0,267	0,342	0,281	0,281	0,133	7	0,267	0,342	0,281
8	0,282	0,327	0,281	0,160	-0,004	8	0,282	0,327	0,281

Enfin, une seconde voie peut-être donnée par une technique de résidus. En effet, de manière générale, quand on redresse un estimateur au moyen d'une technique de calage, on estime la variance au moyen d'une technique de résidus (voir à ce sujet Deville et Samdal 1992, et Deville 1999). Dans le cas du calcul de la variance de Y_p on peut utiliser une technique de résidus pour obtenir la variance exacte. En effet, considérons la variable

$$v^k(l) = \begin{cases} 0 & \text{si } k = r^{l+(h-1)q+1}, \dots, r^{l+hq-1} \\ & \text{si } k = r^{l+(h-1)q} \text{ ou } k = r^{l+hq} \end{cases}$$
$$\left(\frac{N_2(N-n)}{Nn(n-1)} \right)^{-\frac{1}{2}} \left(\frac{N_{h|l}(N_{h|l}-n_{h|l})}{N_{h|l}n_{h|l}} (1 - \frac{Y^k - \bar{Y}_{h|l}}{N_{h|l} - \bar{Y}_{h|l}}) \right)^{\frac{1}{2}}$$

La variable $v^k(l)$ injectée dans l'expression classique du qui peut apparaître comme un résidu associé à l'estimateur Y_p . Cette variable dépend cependant des $\bar{Y}_{h|l}$ qui sont inconnus. On peut estimer $v^k(l)$ par

$$\sum_{k \in U} v^k \left(\frac{N}{N-n} - \frac{1}{N-1} \sum_{k \in U} v^k \right) = V(Y_p|E_l).$$

Si on injecte $v^k(l)$ dans l'estimateur de la variance du plan simple sans remise, on obtient un estimateur sans biais de la variance conditionnelle, et donc de la variance.

$$N^2 \frac{N-n}{N-n} \frac{1}{n} \sum_{j=1}^n \left(v^j - \frac{1}{n} \sum_{l=1}^n v^l \right) = V(Y_l|E_l).$$

Deville (1999) montre que la variance d'une somme de résidus associés à ces estimateurs, les résidus étant calculés par linéarisation. Pour estimer la variance de Y_c , on pourrait donc simplement prendre la moyenne des résidus $v^k(l)$, ce qui s'écrit

$$v^k = \frac{1}{b} \sum_{l=b-1}^b v^k(l).$$

On pourrait ainsi estimer la variance par

$$V_2(Y_c) = \frac{N^2(N-n)}{N-n} \frac{1}{n} \sum_{k \in S} v^k - \left(\frac{1}{n} \sum_{k \in S} v^k \right)^2.$$

9. VARIANCE ET ESTIMATION DE VARIANCE

Pour calculer la variance de \hat{Y}_c , on commence par calculer la variance de \hat{Y}_f . Comme \hat{Y}_f est sans biais conditionnellement à E_f on a

$$V(\hat{Y}_f | E_f) = E V(\hat{Y}_f | E_f).$$

Comme dans chacune des post-strates, conditionnellement à E_f le plan est simple sans remise de taille fixe, on a

$$V(\hat{Y}_f | E_f) = \sum_{h=1}^H N^2 V(\hat{y}_{h|f})$$

$$= \sum_{h=1}^H N^2 \frac{N_{h|f}}{N_{h|f} - n_{h|f}} \frac{n_{h|f}}{S_2^2}$$

(3)

où

$$n_{0|f} = l - 1,$$

$$n_{h|f} = g - 1, h = 1, \dots, H,$$

$$n_{H+1|f} = n - (l + Hg),$$

$$\bar{y}_{0|f} = \frac{1}{l} \sum_{k=1}^l Y^{(k)},$$

$$\bar{y}_{h|f} = \frac{1}{g} \sum_{k=f_l+(h-1)g+1}^{f_l+hg+1} Y^{(k)}, h = 1, \dots, H,$$

$$\bar{Y}_{H+1|f} = \frac{1}{N} \sum_{k=N-f_l+H+1}^N Y^{(k)},$$

$$S_2^2 = \frac{1}{l} \sum_{k=1}^l (Y^{(k)} - \bar{y}_{0|f})^2,$$

$$S_2^2 = \frac{1}{g} \sum_{k=f_l+(h-1)g+1}^{f_l+hg+1} (Y^{(k)} - \bar{y}_{h|f})^2, h = 1, \dots, H,$$

et

$$S_2^2 = \frac{1}{N} \sum_{k=N-f_l+H+1}^N (Y^{(k)} - \bar{Y}_{H+1|f})^2,$$

où les $Y^{(k)}$ représentent les valeurs de Y triées selon l'ordre croissant des $X^{(k)}$.

On remarque qu'il est très difficile de calculer la variance non-conditionnelle de \hat{Y}_f , c'est-à-dire l'espérance de $V(\hat{Y}_f | E_f)$. En effet, $N_{h|f}$ et S_2^2 sont aléatoires. Cependant, on peut estimer simplement $V(\hat{Y}_f | E_f)$ et obtenir un estimateur sans biais de la variance conditionnelle (et donc de la variance) en estimant simplement (3), par

$$V(\hat{Y}_f | E_f) = \sum_{h=1}^H N^2 \frac{N_{h|f}}{N_{h|f} - n_{h|f}} \frac{n_{h|f}}{S_{2|f}^2} \quad (4)$$

où

ce qui revient à estimer l'écart-type des moyennes par la moyenne des écart-types.

ce qui peut être estimé par

$$V(\hat{Y}_c) \leq \frac{1}{b} \sum_{b+g-1}^b \frac{1}{g} \sum_{b+g-1}^b V(\hat{Y}_f) = \frac{1}{b} \sum_{b+g-1}^b \left(\frac{1}{g} \sum_{b+g-1}^b V(\hat{Y}_f) \right)^2,$$

on a un majorant donné par

$$\text{Cov}(\hat{Y}_f, \hat{Y}_f) \leq \sqrt{V(\hat{Y}_f) V(\hat{Y}_f)},$$

variance, comme

Une première voie consiste à chercher un majorant de la

cable, il faut donc chercher une approximation.

Le calcul $E(\hat{Y}_f \hat{Y}_f | E_f)$ semble malheureusement inextric

$$= E E(\hat{Y}_f \hat{Y}_f | E_f) - Y^2.$$

$$\text{Cov}(\hat{Y}_f, \hat{Y}_f) = E \text{Cov}(\hat{Y}_f, \hat{Y}_f | E_f)$$

Comme $E(\hat{Y}_f | E_f) = Y$, obtient

$$+ \text{Cov}(E(\hat{Y}_f | E_f), E(\hat{Y}_f | E_f)).$$

$$\text{Cov}(\hat{Y}_f, \hat{Y}_f) = E \text{Cov}(\hat{Y}_f, \hat{Y}_f | E_f)$$

se fait sans difficulté. Quand $l \neq 1$, il faut calculer

$$V(\hat{Y}_c) = \frac{1}{b} \sum_{b+g-1}^b \frac{1}{g} \sum_{b+g-1}^b \text{Cov}(\hat{Y}_f, \hat{Y}_f).$$

Malheureusement, le calcul de la variance de \hat{Y}_c devient plus complexe à cause des covariances. En effet, on a

LA VARIANCE

10. APPROXIMATIONS POUR LE CALCUL DE

L'estimateur $V(\hat{Y}_f | E_f)$ mais aussi pour $V(\hat{Y}_f)$.

$$S_2^{H+1|f} = \frac{1}{n} \sum_{n=f-f_l+H+1}^{f-f_l+H+1} (Y_f - \bar{y}_{H+1|f})^2.$$

et

$$S_2^{h|f} = \frac{1}{g} \sum_{k=f_l+(h-1)g+1}^{f_l+hg+1} (Y^{(k)} - \bar{y}_{h|f})^2, h = 1, \dots, H,$$

$$S_2^{0|f} = \frac{1}{l} \sum_{k=1}^l (Y_f - \bar{y}_{0|f})^2,$$

8. COMPARAISON À L'ESTIMATEUR PAR LA RÉGRESSION

gain de précision de l'estimateur calé sur la répartition est très important sur la moyenne comme sur les quantiles. L'estimateur que nous proposons se comporte donc de manière robuste en cas de relation non-linéaire entre la variable auxiliaire et la variable d'intérêt.

Afin de comparer les qualités de l'estimateur proposé, un ensemble de simulations a été réalisé pour comparer l'estimateur calé sur la répartition avec l'estimateur de Horvitz-Thompson et l'estimateur par la régression. Trois populations de taille 1 000 ont été générées selon les modèles suivants.

- **Modèle A Dépendance linéaire** : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = X_k + 1,3333 \times \epsilon_k$ où $\epsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $p = 0,616154$.

- **Modèle B Dépendance non linéaire 1** : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = (0,2 + X_k)^2 + 1,3333 \times \epsilon_k$ où $\epsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $p = 0,28975$.

- **Modèle C Dépendance non linéaire 2** : La population est générée selon le modèle $X_k \sim N(0, 1)$ et $Y_k = (0,4 + X_k)^2 + 1,3333 \times \epsilon_k$ où $\epsilon_k \sim N(0, 1)$. Le coefficient de corrélation obtenu dans la population est $p = 0,476158$.

Dans chaque population 100 000 échantillons de taille 100 ont été sélectionnés. Pour chaque échantillon trois systèmes de pondération ont été calculés.

1. les poids associés au plan simple $w_k = N/n$,
2. les poids de l'estimateur calé sur la répartition donné en (2) en prenant la fenêtre $q = 10$ et la bordure $b = 6$,
3. les poids de l'estimateur par la régression donnés par

$$w_k = \frac{n}{N} + \left(X - X_{HT} \right) \left(X_k - \bar{X} \right)^{-\frac{k-5}{k-2}},$$

où X est le total des X_k sur la population, X_{HT} est l'estimateur de Horvitz-Thompson de X_k , et $\bar{X} = X_{HT}/N$.

Au moyen de ces poids, l'estimateur de la moyenne et des neuf déciles ont été calculés pour chaque échantillon. Ensuite, on estime la variance de ces estimateurs au moyen des simulations. Les résultats sont donnés dans les tableaux 2, 3 et 4. Les variances ont été ramenées à 1 pour le plan simple. Pour le modèle linéaire, l'estimateur par la régression est légèrement préférable. Cependant, pour le cas non-linéaire, le

Paramètre	Calage répartition	Estim. Régression
Modèle A : variance des estimateurs (référence : Horvitz-Thompson=1)		
moyenne	0,674422	0,632608
1 ^{er} décile	0,905273	0,893876
2 ^{ème} décile	0,815403	0,802082
3 ^{ème} décile	0,842681	0,815071
4 ^{ème} décile	0,806749	0,768283
5 ^{ème} décile	0,783731	0,740765
6 ^{ème} décile	0,818051	0,782549
7 ^{ème} décile	0,794411	0,773794
8 ^{ème} décile	0,857114	0,844874
9 ^{ème} décile	0,884424	0,884032

Tableau 2

Paramètre	Calage répartition	Estim. Régression
Modèle B : variance des estimateurs (référence : Horvitz-Thompson=1)		
moyenne	0,429689	0,953025
1 ^{er} décile	0,913598	0,958656
2 ^{ème} décile	0,919394	1,009270
3 ^{ème} décile	0,829860	0,987950
4 ^{ème} décile	0,792094	0,989114
5 ^{ème} décile	0,703908	0,992023
6 ^{ème} décile	0,622705	1,009830
7 ^{ème} décile	0,550028	0,981249
8 ^{ème} décile	0,443828	1,010340
9 ^{ème} décile	0,549615	1,029120

Tableau 3

Paramètre	Calage répartition	Estim. Régression
Modèle C : variance des estimateurs (référence : Horvitz-Thompson=1)		
moyenne	0,30768	0,808114
1 ^{er} décile	0,95560	0,983582
2 ^{ème} décile	0,85920	0,970913
3 ^{ème} décile	0,73854	0,930401
4 ^{ème} décile	0,65728	0,950651
5 ^{ème} décile	0,60500	0,956807
6 ^{ème} décile	0,52139	0,930514
7 ^{ème} décile	0,45709	0,907537
8 ^{ème} décile	0,40752	0,903593
9 ^{ème} décile	0,39820	0,860050

Tableau 4

Les poids s'obtiennent simplement en faisant la moyenne des poids des estimateurs Y_0 et Y_1 , et valent

$$w_3 = (6 + 7/2)/2 = 19/4,$$

$$w_7 = (1 + 7/2)/2 = 9/4,$$

$$w_8 = (3 + 1)/2 = 2,$$

$$w_{11} = (1 + 3)/2 = 2,$$

$$w_{12} = (3 + 1)/2 = 2,$$

$$w_{15} = (1 + 4)/2 = 5/2,$$

$$w_{17} = (5 + 4)/2 = 9/2.$$

L'estimateur Y_c est linéaire et strictement sans biais.

7. APPLICATION À L'ESTIMATION DE LA RÉPARTITION

Il existe de multiples méthodes permettant d'utiliser de manière appropriée de l'information auxiliaire pour estimer une fonction de répartition. On peut trouver une exposé de ces techniques dans Ren (2000) et dans Wu et Sitter (2001). La méthode que nous proposons permet également d'estimer la répartition. La répartition dans la population est définie par

$$F_1(y) = \frac{1}{N} \sum_{k \in U} I\{y_k \leq y\},$$

et peut être estimée par

$$\hat{F}_1(y) = \frac{\sum_{k \in S} w_k I\{y_k \leq y\}}{\sum_{k \in S} w_k},$$

où $I\{y \leq y_k\}$ est la fonction indicatrice, et les w_k sont les poids associés aux unités k qui valent $1/\pi_k = N/n$ pour l'estimateur de Horvitz-Thompson, et qui sont donnés en

(2) pour l'estimateur cale.

Notons que les deux fonctions sont discrètes, mais que les sauts sont beaucoup moins nombreux sur S que sur U . Pour atténuer les différences des répartition entre l'échantillon et la population, nous avons lissé les fonctions de répartition, en utilisant comme Deville (1995) une interpolation linéaire des centres des contremaiches, ce qui consiste à définir $F_2(y)$ en reliant les points

$$\frac{1}{2} \{F_1(y_k) - F_1(y_k - \varepsilon)\},$$

pour $k \in U$, où ε est un réel strictement positif arbitrairement petit. Ensuite, on définit $F_2(y)$ en reliant les points

$$\frac{1}{2} \{F_1(y_k) - F_1(y_k - \varepsilon)\},$$

pour l'échantillon.

Exemple 2 : Une population de taille $N = 1\,000$ a été générée au moyen de variables logarithmico-normales

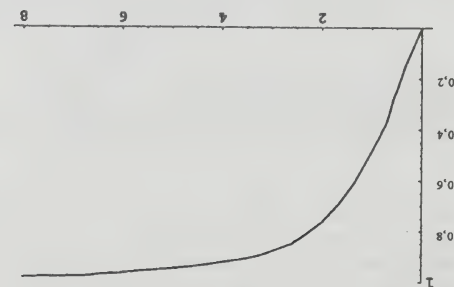


Figure 1. Fonction de répartition dans la population

Ensuite, la figure 2 montre $F_2(x)$ et la répartition estimée par l'estimateur de Horvitz-Thompson. Enfin, la figure 3 montre $F_2(x)$ et la répartition estimée par l'estimateur cale.

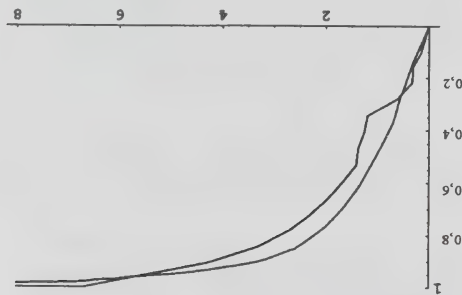


Figure 2. Fonction de répartition dans la population et estimateur de Horvitz-Thompson de la répartition

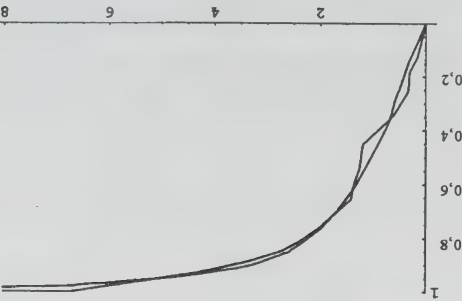


Figure 3. Fonction de répartition dans la population et estimateur cale de la répartition

indépendantes et équidistribuées. Un échantillon de taille $n = 16$ a ensuite été sélectionné et on a fixé $h = 5$. La figure 1 donne $F_2(x)$ dans la population.

$$w_j = \frac{1}{1} \left(\sum_{b=0}^{l-b} \frac{q}{r_{j+l-b} - r_{j+l-b-q} - 1} q - 1 \right) = \frac{1}{b-1} \sum_{a=0}^{b-1} (r_{j+a} - r_{j+a-q}).$$

Si toutes les valeurs de la variable auxiliaire ne sont pas distinctes, on peut affecter les rangs unités ayant des valeurs communes au hasard. Par exemple, si on a $X_1^* = 2, X_2^* = 5, X_3^* = 5, X_4^* = 7, X_5^* = 8$, on choisit avec une probabilité $1/2$, soit les rangs $R_1^* = 1, R_2^* = 2, R_3^* = 3, R_4^* = 4, R_5^* = 5$, soit $R_1^* = 1, R_2^* = 3, R_3^* = 2, R_4^* = 4, R_5^* = 5$. On calcule alors l'estimateur hissé pour chaque permutation, et on fait leur moyenne. Cette méthode a l'avantage de conserver un estimateur sans biais. En effet, pour chacune des permutations possibles, l'estimateur est sans biais. En pratique, il est n'est pas nécessaire de calculer les estimateurs pour toutes les permutations. On peut calculer l'estimateur d'une permutation, et ensuite, on égalise simplement les poids des unités ayant les mêmes valeurs pour la variable x .

6. CAS où $q = 2, b = 2$

Quand $q = 2$, et $b = 2$, on obtient après quelques calculs

$$\begin{aligned} \bar{y}^c = \frac{1}{2} \left\{ \sum_{j=3}^{f-2} y_j (r_{j+1} - r_{j-1}) + \frac{r_{j+1} - r_{j-2} + 1}{2} y_{n-1} + \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n \right\} \\ + \frac{r_{j+2} - r_{j-3}}{2} y_1 + \frac{r_{j+1} - r_{j-3}}{2} y_2 + \frac{r_{j+1} - r_{j-2} + 1 - 2r_n}{2} y_{n-1} + y_{n-1} \frac{2}{r_{n+1} - r_{n-2} - 3} \end{aligned}$$

où $r_0 = 0$ et $r_{n+1} = N + 1$. On retrouve un estimateur proposé par Ren (2000, page 140) et obtenu avec un argument de calage. Seule la gestion des bords est légèrement différente. **Exemple 1** : Soit une population de taille $N = 20$. Supposons que les valeurs de la variable d'intérêt se trouvent dans la Table 1. On suppose, en outre que l'échantillon de taille $n = 7$ est composé des unités de rangs $\{3, 7, 8, 11, 12, 15, 17\}$. Si on prend $q = 2, l = 2, b = 2$ on obtient $E_2 = \{r_2, r_4, r_6\} = \{7, 11, 15\}$. Ensuite on peut calculer $E(r_k | E_2) = \{7, 11, 15\}$. Les probabilités d'inclusion conditionnelles sont les suivantes :

$$\begin{aligned} E(I_3 | E_2) = \{7, 11, 15\} &= 1/6, \\ E(I_7 | E_2) = \{7, 11, 15\} &= 1, \\ E(I_8 | E_2) = \{7, 11, 15\} &= 1/3, \\ E(I_{11} | E_2) = \{7, 11, 15\} &= 1, \\ E(I_{12} | E_2) = \{7, 11, 15\} &= 1/3, \\ E(I_{15} | E_2) = \{7, 11, 15\} &= 1, \\ E(I_{17} | E_2) = \{7, 11, 15\} &= 1/5. \end{aligned}$$

Tableau 1

Exemple d'une population de taille $N = 20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	71	72	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	13	8	17	9	16	12	19	4	5	11	18	10

L'estimateur

$$\bar{y}_0^c = \sum_{j_k} \frac{E(I_k | E_2 = \{7, 11, 15\})}{y_k}$$

est donc sans biais et conditionnellement sans biais. De plus, il est linéaire et

$$\sum_{k \in S} \frac{E(I_k | E_2 = \{7, 11, 15\})}{1} = N.$$

D'autre part, si on prend $q = 2, l = 3, b = 2$, on obtient $E_3 = \{r_3, r_5, r_7\} = \{8, 12\}$. Alors, avec le même exemple, on calcule $E(I_k | E_3) = \{8, 12\}$, et on obtient

$$\begin{aligned} E(I_3 | E_3) = \{8, 12\} &= 2/7, \\ E(I_7 | E_3) = \{8, 12\} &= 2/7, \\ E(I_8 | E_3) = \{8, 12\} &= 1, \\ E(I_{11} | E_3) = \{8, 12\} &= 1/3, \\ E(I_{12} | E_3) = \{8, 12\} &= 1, \\ E(I_{15} | E_3) = \{8, 12\} &= 2/8 = 1/4, \\ E(I_{17} | E_3) = \{8, 12\} &= 2/8 = 1/4. \end{aligned}$$

L'estimateur

$$\bar{y}_1^c = \sum_{j_k} \frac{E(I_k | E_3 = \{8, 12\})}{y_k}$$

est également sans biais et linéaire. Enfin, on calcule la moyenne des deux estimateurs :

$$\bar{y}^c = \frac{\bar{y}_0^c + \bar{y}_1^c}{2}.$$

Horvitz-Thompson dans les plans simples, n'est donc pas perdue. Les unités dont les rangs sont dans E_l^j sont appelées unités pivots, et sont affectées d'un poids égal à 1, ce qui rend les poids très inégaux. On peut donc reprocher à \hat{Y}_l^j d'utiliser des poids fortement dispersés. Ce problème peut être résolu en réalisant un lissage des estimateurs.

5. LISSAGE DES ESTIMATEURS

Pour résoudre le problème de la dispersion des poids, on réalise une moyenne mobile d'estimateurs de la manière suivante :

$$\hat{Y}_l^c = \frac{1}{b} \sum_{l=b+1}^l \hat{Y}_l^j$$

\hat{Y}_l^c garde toutes les propriétés des \hat{Y}_l^j . Il est donc sans biais, calé sur N et linéaire, il peut donc s'écrire sous la forme

$$\hat{Y}_l^c = \sum_{j=1}^f w_j \hat{Y}_l^j,$$

où $w_j =$

$$\left\{ \begin{array}{l} \frac{1}{b} \sum_{l=b+1}^l \frac{b}{f_l - 1} \\ \frac{1}{b} \sum_{l=b+1}^l \frac{b}{f_l - 1} \frac{f_l - l - 1 - m}{f_l - l - 1 - m - f_l - l - 1 - b - q - 1} + 1, \quad b \leq j < b + 1, \\ \frac{1}{b} \sum_{l=b+1}^l \frac{b}{f_l - 1} \frac{f_l - l - 1 - m}{f_l - l - 1 - m - f_l - l - 1 - b - q - 1} + 1, \quad b + 1 \leq j \leq u - b + 2 - q, \\ \frac{1}{b} \sum_{l=b+1}^l \frac{b}{f_l - 1} \frac{f_l - l - 1 - m}{f_l - l - 1 - m - f_l - l - 1 - b - q - 1} + 1, \quad u - b + 2 - q < j \leq u - b + 1, \\ \frac{1}{b} \sum_{l=b+1}^l \frac{b}{f_l - 1} \frac{f_l - l - 1 - m}{f_l - l - 1 - m - f_l - l - 1 - b - q - 1} + 1, \quad u - b + 1 < j, \end{array} \right.$$

(2)

$$m^-(x) = \begin{cases} 0 & \text{si } x > b \\ x & \text{sinon,} \end{cases} \quad m^+(x) = \begin{cases} u + 1 & \text{si } x > u - b + 1 \\ x & \text{sinon,} \end{cases}$$

$$r_0^+ = 0 \text{ et } r_{u+1}^- = N + 1.$$

Sous l'apparente complexité due au traitement particulier des bords, le système de pondération est relativement simple. Dans le cas où l'on ne se trouve pas trop près des bords, il vaut alors

$$= \frac{1}{r_l^+ - 1} \sum_{j=1}^f \hat{Y}_l^j + \hat{Y}_l^j$$

$$+ \left(\sum_{h=1}^H \frac{1}{r^{l+Hq} - r^{l+(H-1)q} + \hat{Y}_l^{l+Hq}} \right) \sum_{j=1}^f \hat{Y}_l^{l+(H-1)q} + \hat{Y}_l^{l+Hq}$$

$$+ \frac{N - r^{l+Hq}}{\sum_{j=1}^f \hat{Y}_l^{l+Hq} + \hat{Y}_l^{l+Hq}}$$

$$= N_{0|l} \hat{Y}_{0|l} + \sum_{h=1}^H \left(N_{h|l} \hat{Y}_{h|l} + \hat{Y}_l^{l+Hq} \right) + N_{H+1|l} \hat{Y}_{H+1|l}$$

$$N_{0|l} = r_l^+ - 1,$$

$$N_{h|l} = r^{l+Hq} - r^{l+(H-1)q}, \quad h = 1, \dots, H,$$

$$N_{H+1|l} = N - r^{l+Hq},$$

$$\hat{Y}_{0|l} = \frac{1}{\sum_{j=1}^f \hat{Y}_l^j} \frac{1}{b} \sum_{j=1}^f \hat{Y}_l^j,$$

$$\hat{Y}_{h|l} = \frac{1}{\sum_{j=1}^f \hat{Y}_l^j} \frac{1}{b} \sum_{j=1}^f \hat{Y}_l^j \frac{1}{r^{l+Hq} - r^{l+(H-1)q}}, \quad h = 1, \dots, H,$$

$$\hat{Y}_{H+1|l} = \frac{1}{\sum_{j=1}^f \hat{Y}_l^j} \frac{1}{b} \sum_{j=1}^f \hat{Y}_l^j \frac{1}{r^{l+Hq} - r^{l+(H-1)q}}, \quad h = 1, \dots, H,$$

Cet estimateur est en réalité un estimateur post-stratifié où les tailles des post-strates sont fixées dans l'échantillon. Comme $E_l^j(E_l^j) > 0$, \hat{Y}_l^j est strictement sans biais non conditionnellement et conditionnellement à E_l^j , ce qui n'est évidemment pas le cas de l'estimateur post-stratifié classique, car ce dernier a une probabilité non-nulle d'avoir une post-strate vide. En fixant la taille des post-strates dans l'échantillon, la constitution de post-strates vides devient impossible. La taille correspondante de la post-strate dans la population est une variable aléatoire $N_{h|l}^j$. L'estimateur \hat{Y}_l^j possède une autre propriété intéressante. En utilisant la définition des $E_l^j(E_l^j)$, on montre assez facilement que

$$\sum_{k \in S} \frac{E_l^k(E_l^k)}{1} = N.$$

L'estimateur est donc calé sur la taille de la population. Cette propriété, que possède également l'estimateur de

— un entier l tel que $b \leq l \leq b + q - 1$, définissant le décalage.

Les quantités q, b , et l servent à définir un sous-ensemble d'indices :

$$E_l = \{r_{l+q}, r_{l+2q}, \dots, r_{l+hq}, \dots, r_{l+q}\},$$

$$\text{pour } l = b, \dots, b + q - 1.$$

Par exemple, si $n = 18, q = 4, b = 3$, alors

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\},$$

$$E_4 = \{r_4, r_8, r_{12}, r_{16}\},$$

$$E_5 = \{r_5, r_9, r_{13}\},$$

$$E_6 = \{r_6, r_{10}, r_{14}\}.$$

La probabilité d'inclusion conditionnelle est calculée par

rapport à l'un des E_l . La valeur de H est définie de manière à ce que

$l + Hq \leq n - b + 1$ et donc H est le plus grand entier tel que $H \leq (n - b - l + 1)/q$. Il est donc clair que H dépend de l .

Ensuite, on peut calculer les probabilités d'inclusion :

$$E(l^k | E_l) = \begin{cases} 1 & \text{si } k \in E_l \\ \frac{q-1}{r_{l+(h-1)q} - 1} & \text{si } r_{l+(h-1)q} < k < r_{l+hq}, h = 1, \dots, H \\ \frac{l-1}{r_l - 1} & \text{si } k < r_l \\ \frac{n - (l + Hq)}{n - r_{l+Hq}} & \text{si } k > r_{l+Hq} \end{cases}$$

Ces probabilités d'inclusion sont donc assez contrastées. Cependant elles sont toutes positives, y compris sur les bords. Il est important de prendre une bordure $b \geq 2$ afin que la première et la dernière post-strate ne soit pas vide.

4. UNE CLASSE D'ESTIMATEURS SANS BIAIS

Comme $E(l^k | E_l) > 0$, on peut construire un estimateur qui est sans biais et même conditionnellement sans biais par rapport à E_l . En notant $y_1, \dots, y_j, \dots, y_n$ les n valeurs prises par les unités dans l'échantillon ordonnées selon les R_k , on obtient

Une des possibilités consiste à prendre $w_k = 1/\pi_k, n/N$, ce qui donne l'estimateur de Horvitz-Thompson,

$$Y_{HT}^w = \sum_{k \in S} \frac{\pi_k}{Y_k} = \frac{n}{N} \sum_{k \in S} Y_k,$$

qui est sans biais.

Nous allons cependant nous intéresser à la classe plus générale des estimateurs pondérés conditionnellement (Tillé 1998, 1999a) où les unités sont pondérées par des inverses de probabilités d'inclusion conditionnelles. Si Z est une statistique quelconque, alors l'estimateur pondéré conditionnellement

$$Y_Z^w = \sum_{k \in S} \frac{Y_k}{Z_k} E(l^k | Z) \quad (1)$$

est strictement sans biais si et seulement si $E(l^k | Z) > 0$, pour tout $k \in U$. En effet,

$$E(Y | Z) = \sum_{k \in U} \frac{E(l^k | Z) Y_k}{E(l^k | Z) Y_k} = Y.$$

Comme l'estimateur est conditionnellement sans biais, il est également non-conditionnellement sans biais. L'estimateur (1) généralise, selon le choix de la statistique Z utilisée, l'estimateur stratifié, mais aussi (à une approximation près) l'estimateur par la régression (voir Tillé 1998).

3. CONDITIONNEMENT SUR DES RANGS

Supposons maintenant que les N valeurs X_1, \dots, X_N d'un caractère auxiliaire x soient connues sur les N unités de la population. Dans un premier temps, on suppose que tous les X_k prennent des valeurs distinctes, cette hypothèse sera ensuite levée en section 5. Le rang R_k de l'unité k est

$$R_k = \#\{l \in U | X_l \leq X_k\}.$$

On note par ailleurs $r_j, j = 1, \dots, n$, les rangs de la population ordonnés des n unités sélectionnées dans l'échantillon donc $r_1 < r_2 < \dots < r_{n-1} < r_n$. Les r_j sont des variables aléatoires ayant une distribution hypergéométrique négative (voir Tillé 1999b).

La statistique utilisée pour définir les probabilités d'inclusion conditionnelles est un sous-ensemble de $\{r_1, \dots, r_j, \dots, r_n\}$. On définit d'abord

- un entier q tel que $2 \leq q \leq n$, définissant la période,
- un entier b tel que $2 \leq b$, définissant la bordure,

Estimation sans biais par calage sur la répartition dans les plans simples sans remise

YVES TILLÉ¹

RÉSUMÉ

L'estimateur post-stratifié a parfois des post-strates vides. Pour pallier ce problème, on construit un estimateur post-stratifié dont les tailles des post-strates sont fixées dans l'échantillon. Les tailles des post-strates sont alors aléatoires dans la population. Ensuite, on construit un estimateur mobile des estimateurs post-stratifiés. Cette technique permet de construire une théorie exacte du calage sur la répartition. L'estimateur obtenu est non seulement calé sur la répartition, il est linéaire, et exactement sans biais. On compare ensuite l'estimateur calé à l'estimateur par la régression. On propose enfin un estimateur approché de la variance valide au moyen de simulations.

MOTS CLÉS : Estimation sans biais; calage sur une fonction de répartition; probabilités d'inclusion conditionnelles; pondération.

1. INTRODUCTION

Lors d'une enquête par sondage, il arrive que l'on connaisse les valeurs d'un caractère auxiliaire pour toutes les unités de la population. Cette information peut être disponible quand les unités sont sélectionnées dans une base de données qui contient d'autres variables d'intérêt. On est alors tenté de caler les résultats d'une enquête sur cette information auxiliaire. Soit, on ne retient de cette variable auxiliaire que certaines fonctions (moments, effectifs) en vue d'utiliser une méthode de calage (voir par exemple Deville et Särndal 1992 ou Estévaço, Hidroglou et Särndal 1995), soit on peut découper cette variable en classes en vue d'utiliser un estimateur post-stratifié.

Si l'on opte pour l'estimateur post-stratifié, le découpage en strates s'avère délicat. Théoriquement, les strates doivent être définies avant la sélection de l'échantillon. Or faut-il placer les bornes des post-strates ? De quelles tailles doivent être les post-strates ? Cette dernière question est la plus embarrassante, car le problème principal de la post-stratification est la possibilité d'obtenir des post-strates vides. Les tailles des post-strates doivent donc être suffisamment grandes pour que la probabilité d'obtenir une post-strate vide soit négligeable. Ces problèmes ne se limitent pas aux estimateurs post-stratifiés. En effet, les estimateurs par la régression ou calés peuvent également ne pas exister pour certains échantillons.

Notre objectif est de définir une nouvelle méthode permettant d'utiliser l'information auxiliaire dans la population. Cette méthode est basée sur la définition de post-strates pour lesquelles le nombre d'unités est fixé dans l'échantillon, et non dans la population. On peut ainsi importer dans l'estimateur une information auxiliaire complexe issue de la connaissance de toutes les valeurs prises par la variable auxiliaire, tout en évitant à la fois le

2. NOTATION

On suppose que l'on a une population composée de N unités d'observation dont les étiquettes des observations sont notées $\{1, \dots, k, \dots, N\}$. Dans cette population, on s'intéresse à un caractère d'intérêt $Y_k, k \in U$. L'objectif consiste à estimer le total $Y = \sum_{k \in U} Y_k$. On sélectionne un échantillon aléatoire S de taille fixe n au moyen d'un plan aléatoire simple sans remise. On note I_k la variable aléatoire indicatrice qui prend la valeur 1 si l'unité k est dans l'échantillon et 0 sinon. Les probabilités d'inclusion dans l'échantillon sont donc définies par $\Pr(k \in S) = \pi_k = n/N, k \in U$, et les probabilités d'inclusion d'ordre deux par $\Pr(k, l \in S) = \pi_{kl} = n(n-1)/(N(N-1)), k \neq l \in U$.

On s'intéressera à la classe des estimateurs linéaires de Y qui s'écrit

problème de la définition des bornes des post-strates et le problème des post-strates vides.

Cet article est organisé comme suit : En section 2, la notation est définie, en section 3, on donne le principe du conditionnement par les rangs, qui permet de définir des estimateurs sans biais en section 4. Ensuite, en section 5, on définit l'estimateur lissé, et un cas particulier est examiné en détail en section 6. En section 7, on donne une application à l'estimation d'une fonction de répartition. En section 8, on compare ce nouvel estimateur à l'estimateur par la régression et l'estimateur du plan simple sans remise. Le calcul de variance est abordé en section 9. Suite à l'impossibilité de donner une solution exacte, on donne une approximation en section 10, qui est testée par des simulations en section 11. Enfin, des conclusions générales sont données en section 12.

Obtention de (3.8), autre estimation de la variance de $\hat{\mu}_{psw}$:

Puisque

$$\frac{N_{gh}}{N_{gh}^{-1} \sum_{k \in s_{gh}} p_k^{-1}} \approx \frac{N_{gh}}{N_{gh}^{-1} n_{gh}} \approx \frac{N_{gh}}{N_{gh}^{-1} n_{gh}} \approx \frac{N_{gh}}{N_{gh}^{-1} \sum_{k \in s_{gh}} p_k^{-1}},$$

le résultat découle de (3.7).

BIBLIOGRAPHIE

- BRICK, J.M., WAKSBERG, J. et KEETER, S. (1994). Evaluating the use of data on interruptions in telephone service for nontelephone households. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 19-28.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.
- KHURSHID, A., et SAHAL, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- MCCULLAGH, P., et NELDER, J.A. (1991). *Generalized Linear Models*. New York: Chapman and Hall.
- RAO, J.N.K. (1997). Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics*, 25, 1-21.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- SÄRNDA, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- STEEH, C.G., GROVES, R.M., COMMENT, R. et HANSMIRE, E. (1983). Report on the survey research center's surveys of consumer attitudes. *Incomplete Data in Sample Surveys*, (Ed. W.G. Madow, H. Niselson et I. Olkin), Academic Press, New York, 1.
- THORNBERY, JR., O.T., et MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Surveys*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, et J. Waksberg). New York: John Wiley & Sons, Inc., 25-49.

lorsque $n \rightarrow \infty$. En outre, puisque

$$E(\tilde{a}^{ps(1)} - \mu_{T,ps} \tilde{a}^{ps(2)} | U^T, n^{gh}) \\ = N^{-1} \sum_{g=1}^G \sum_{h=1}^H N^{gh} N^{Tgh} \sum_{k \in U_{T,ps}^{gh}} (y_{1k} - \mu_{T,ps} y_{2k}), \quad (A.8)$$

$$E[\text{var} E(\tilde{a}^{ps(1)} - \mu_{T,ps} \tilde{a}^{ps(2)} | U^T, n^{gh}) | U^T] = 0. \quad (A.9)$$

Puisque le théorème A.1 et (A.7) impliquent

$$\text{var} \hat{\mu}^{ps} = (\alpha_2^*)^{-2} E \text{var} (\tilde{a}^{ps(1)} - \mu_{T,ps} \tilde{a}^{ps(2)} | U^T, n^{gh}) \\ + O(n^{-2} + N^{-1}) \quad (A.10)$$

lorsque $n \rightarrow \infty$, alors (A.9) implique (4.3).

Obtention de (3.1), la variance estimative de $\hat{\mu}^{ps}$:

Étant donné (A.6), nous avons l'estimateur

$$\widehat{\text{var}} \left(n_{gh}^{-1} \sum_{k \in s_{gh}} \{y_{1k} - \mu_{T,ps} y_{2k}\} | U^T, n^{gh} \right) \\ = \frac{1 - n_{gh}^{gh}/N^{Tgh}}{\sum_{k \in s_{gh}} n_{gh}^{gh} (n_{gh}^{gh} - 1)} =$$

$$\left[y_{1k} - \mu_{T,ps} y_{2k} - n_{gh}^{gh} (y_{1f} - \mu_{T,ps} y_{2f}) \right]^2 \sum_{j \in s_{gh}}^{-1}$$

Le résultat découle de l'utilisation de (A.10).

Obtention de (3.2), biais estimatif de $\hat{\mu}^{ps}$:

Le lemme A.1 implique

$$\hat{\mu}^{ps} - \mu^* = O(n^{-1})$$

lorsque $n \rightarrow \infty$. Puisque

$$E \tilde{a}^{ps(i)} = \alpha_i^* + O(N^{-1})$$

lorsque $N \rightarrow \infty$ pour $i = 1, 2$, le résultat s'ensuit.

A.2 Estimateur pondéré en fonction de l'abonnement au téléphone

Ici nous obtenons les équations liées à l'estimateur pondéré en fonction de l'abonnement au téléphone, $\hat{\mu}^{w,}$ sous les conditions C et D, où $\tilde{a}^{w(1)}$ et $\tilde{a}^{w(2)}$ satisfont les conditions C et D. Notons que

$$E(\tilde{a}^{w(i)} | U^T) = N^{-1} \sum_{k \in U^T} y_{1k} / p_k \\ \text{pour } i = 1, 2, \text{ et définitions } \mu_{T,w} = E(\tilde{a}^{w(1)} | U^T) / E(\tilde{a}^{w(2)} | U^T).$$

implique

$$E(\tilde{a}^{psw(i)} | U^T) = N^{-1} \sum_{g=1}^G \sum_{h=1}^H N^{gh} \sum_{k \in U_{T,gh}} \frac{p_k^{-1} y_{1k}}{p_k^{-1} y_{2k}} \\ + n^{-1} \varepsilon_n$$

a posteriori, $\hat{\mu}^{psw,}$ sous les conditions C et D, où $\tilde{a}^{psw(1)}$ et $\tilde{a}^{psw(2)}$ satisfont les conditions C et D. Le lemme A.1

l'abonnement au téléphone

A.3 Estimateur stratifié a posteriori pondéré en fonction de

Le résultats s'ensuit.

$$(\alpha_2^*)^{-2} E \text{var} \left(N^{-1} \sum_{h=1}^H \frac{N^{Th}}{N^{Th}} \sum_{k \in s_h} \frac{p_k}{y_{1k} - \mu_{T,w} y_{2k}} | U^T \right)$$

qui est équivalent à

$$(\alpha_2^*)^{-2} E \text{var} (\tilde{a}^{w(1)} - \mu_{T,w} \tilde{a}^{w(2)} | U^T),$$

de var $\hat{\mu}^{w,}$ estime aussi

Étant donné le théorème A.1, une estimation valide

Obtention de (3.6), la variance estimative de $\hat{\mu}^{ps}$:

lorsque $n \rightarrow \infty$. Le résultat découle de l'application du théorème A.1 à (A.11).

$$(A.12) \quad + O(n^{-1} N^{-1/2})$$

$$\left[\sum_{j \in U_h} p_k \frac{y_{1k} - \mu_{T,w} y_{2k}}{\sum_{j \in U_h} p_j} - \frac{\sum_{j \in U_h} (y_{1f} - \mu_{T,w} y_{2f})}{\sum_{j \in U_h} p_j} \right]^2$$

$$(A.11) \quad = N^{-2} \sum_{h=1}^H \left[\frac{\sum_{j \in U_h} p_j}{\sum_{j \in U_h} p_j} - n_h \right] \left[\frac{\sum_{j \in U_h} n_h p_j}{\sum_{j \in U_h} p_j} - 1 \right]$$

$$E \text{var} (\tilde{a}^{w(1)} - \mu_{T,w} \tilde{a}^{w(2)} | U^T)$$

alors

$$\left[y_{1k} - \mu_{T,w} y_{2k} - N^{-1} \sum_{j \in U_h} n_h \frac{p_k}{y_{1f} - \mu_{T,w} y_{2f}} \right]^2$$

$$= N^{-2} \sum_{h=1}^H \sum_{k \in U_{T,h}} \frac{n_h (N^{Th} - n_h) N^{Th}}{(N^{Th} - 1)}$$

$$\text{var} (\tilde{a}^{w(1)} - \mu_{T,w} \tilde{a}^{w(2)} | U^T)$$

Puisque

Obtention de (4.4), la variance asymptotique de $\hat{\mu}^{ps}$:

lorsque $k \rightarrow \infty$. Donc,

$$\text{cov}\{(\alpha_{1k} - \mu \alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} = O(k^2) \quad (\text{A.2})$$

lorsque $k \rightarrow \infty$. La première partie du lemme découle de la combinaison de (A.1) et (A.2). Puisque le lemme A.1 implique

$$\text{biais } u_k = O(k^{-1})$$

lorsque $k \rightarrow \infty$, alors la deuxième partie de ce lemme en

découle.

Condition C : Si nous définissons $\alpha_{Ti} = \text{plim}_{n \rightarrow \infty} \hat{\alpha}_i$ étant donné U_T , l'estimateur $\hat{\alpha}_i^p$ de α_i satisfait les conditions suivantes pour $i = 1, 2$:

$$E(\hat{\alpha}_i | U_T) - \alpha_{Ti} = O(n^{-1});$$

$$\text{étant donné } U_T, \hat{\alpha}_i - \alpha_{Ti} = O^p(n^{-1/2});$$

et

$$E(|\hat{\alpha}_i - \alpha_{Ti}|_3 | U_T) = O(n^{-3/2})$$

lorsque $k \rightarrow \infty$.

Condition D : Étant donné U_T

$$n^{1/2} \left[\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} - \begin{pmatrix} \alpha_{T1} \\ \alpha_{T2} \end{pmatrix} \right] \xrightarrow{d} \bar{d}$$

$$N \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} n^{-1/2} \xi_n \\ n^{-1/2} \xi_n \end{pmatrix}$$

pour une certaine matrice définie positive Σ , où $\alpha_{Ti} = \text{plim}_{n \rightarrow \infty} \hat{\alpha}_i$ étant donné U_T . En outre,

$$E(|\hat{\alpha}_i - \alpha_{Ti}|_3 | U_T) = O(n^{-3/2})$$

lorsque $n \rightarrow \infty$, pour $i = 1, 2$.

THÉOREME A.1 Si les conditions C et D sont remplies, nous avons

$$\text{var } \hat{\mu} = \alpha_{T2}^{-2} E \text{ var}(\hat{\alpha}_1 - \mu^T \hat{\alpha}_2 | U_T) + O(n^{-2} + N^{-1})$$

lorsque $n \rightarrow \infty$, où $\mu^T = \alpha_{T1} / \alpha_{T2}$.

PREUVE : Pour commencer, nous déterminons $E \text{ var}(\hat{\mu} | U_T)$. Sous la condition D, nous appliquons le lemme A.2 pour obtenir

$$\text{var}(\hat{\mu} | U_T) = \alpha_{T2}^{-2} \text{var}(\hat{\alpha}_1 - \mu^T \hat{\alpha}_2 | U_T) + n^{-2} \xi_n. \quad (\text{A.3})$$

Puisque

$$\alpha_{T2}^{-2} = (\alpha_2)^{-2} + N^{-1/2} \xi_n$$

et que

$$\text{var}(\hat{\alpha}_1 - \mu^T \hat{\alpha}_2 | U_T) = n^{-1} \xi_n,$$

alors (A.3) implique

$$E \text{ var}(\hat{\mu} | U_T) =$$

$$\alpha_{T2}^{-2} E \text{ var}(\hat{\alpha}_1 - \mu^T \hat{\alpha}_2 | U_T) + O(n^{-2} + n^{-1} N^{-1/2}) \quad (\text{A.4})$$

lorsque $n \rightarrow \infty$. Maintenant, nous déterminons $\text{var } E(\hat{\mu} | U_T)$. La condition C et le lemme A.1 implique

$$E(\hat{\mu} | U_T) = \mu^T + n^{-1} \xi_n = \mu + (n^{-1} + N^{-1/2}) \xi_n.$$

Donc,

$$\text{var } E(\hat{\mu} | U_T) = O(n^{-2} + N^{-1}) \quad (\text{A.5})$$

lorsque $n \rightarrow \infty$. Le résultat découle de la combinaison de

(A.4) avec (A.5).

A.1 Estimateur stratifié à posteriori

Ici, nous établissons les équations liées à l'estimateur stratifié à posteriori, $\hat{\mu}^{\text{ps}}$, où $\hat{\alpha}^{\text{ps}(1)}$ et $\hat{\alpha}^{\text{ps}(2)}$ satisfont les conditions C et D. Notons que

$$E(\hat{\alpha}^{\text{ps}(1)} | U_T) = N^{-1} \sum_{H=1}^h \sum_{G=1}^g N_{gh}^{-1} \sum_{k \in U_{Tgh}} y_{ik}$$

pour $i = 1, 2$, et définissons $\mu_{T, \text{ps}} = E(\hat{\alpha}^{\text{ps}(1)} | U_T)$. Rappelons les définitions de α_i et μ^* dans (4.1) et (4.2).

Obtention de (4.3), la variance asymptotique de $\hat{\mu}^{\text{ps}}$:

$$\text{var}(\hat{\alpha}^{\text{ps}(1)} | U_T, n^{gh}) = N^{-2} \sum_H \sum_{G=1}^g \frac{N_{gh}^h}{N_{Tgh}^h} \sum_{k \in U_{Tgh}} \left[y_{1k} - \mu_{T, \text{ps}} y_{2k} - N_{Tgh}^{-1} \sum_{j \in U_{Tgh}} (y_{1j} - \mu_{T, \text{ps}} y_{2j}) \right]^2$$

Alors

$$E \text{ var}(\hat{\alpha}^{\text{ps}(1)} | U_T, n^{gh}) = \mu_{T, \text{ps}}^2 \hat{\alpha}^{\text{ps}(2)} | U_T, n^{gh}$$

$$= N^{-2} \sum_H \sum_{G=1}^g \frac{N_{gh}^h}{N_{Tgh}^h} \left[\sum_{j \in U_{Tgh}} d_j \left(y_{1j} - \mu^* y_{2j} \right) \right] \left[\sum_{j \in U_{Tgh}} d_j \left(y_{1j} - \mu^* y_{2j} \right) \right]^2$$

$$\sum_{k \in U_{Tgh}} d_k \left[y_{1k} - \mu^* y_{2k} - \sum_{j \in U_{Tgh}} \frac{d_j}{\sum_{j \in U_{Tgh}} d_j} (y_{1j} - \mu^* y_{2j}) \right]^2$$

$$+ O(n^{-2} + n^{-1} N^{-1/2}) \quad (\text{A.7})$$

(mais en commentant un léger abus), nous permettrons que la série représentée par $\{\xi_1, \xi_2, \dots\}$ diffère d'une équation à l'autre.

Condition A : Chaque α_k représente une moyenne d'observations sur échantillon telle que $E\alpha_k - \alpha_i = O(k^{-1})$, $E|\alpha_k - \alpha_i|^3 = O(k^{-3/2})$ et $\alpha_i = O^p(k^{-1/2})$ lorsque $k \rightarrow \infty$ pour $i = 1, 2$. Posons que $\mu_k = \alpha_1 k / \alpha_2 k$ pour $k = 1, 2, \dots$.

LEMME A.1 La condition A implique $E\mu_k - \mu = O(k^{-1})$ lorsque $k \rightarrow \infty$.

PREUVE : Définissons la fonction $f(\gamma_1, \gamma_2) = \gamma_1 / \gamma_2$. En vertu d'un développement linéaire en série de Taylor,

$$\begin{aligned} \mu_k - \mu &= \alpha_1 k / \alpha_2 k - \alpha_1 / \alpha_2 \\ &= (\alpha_1 k - \alpha_1) (\alpha_2)^{-1} - (\alpha_2 k - \alpha_2) \mu (\alpha_2)^{-2} + k^{-1} \xi_k. \\ &= (\alpha_1 k - \alpha_1) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_2 k - \alpha_2) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} + k^{-1} \xi_k \end{aligned}$$

Le résultat découle de la condition A.

Condition B : La série $\{\alpha_{i1}, \alpha_{i2}, \dots\}$ pour $i = 1, 2$ satisfait

$$k^{1/2} \begin{pmatrix} \alpha_{1k} \\ \alpha_{2k} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \begin{pmatrix} \bar{d} \\ \bar{d} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \sigma_2^2 & \rho \sigma_1 \sigma_2 \end{pmatrix} + \begin{pmatrix} k^{-1/2} \xi_{1k} \\ k^{-1/2} \xi_{2k} \end{pmatrix}$$

pour certaines constantes σ_1^2, σ_2^2 et ρ .

LEMME A.2 Si les conditions A et B sont remplies,

$$E\bar{Q}\mu_k = (\alpha_2)^{-2} \text{var}(\alpha_1 k - \mu \alpha_2 k) + O(k^{-2}),$$

et

$$\text{var} \mu_k = (\alpha_2)^{-2} \text{var}(\alpha_1 k - \mu \alpha_2 k) + O(k^{-2}),$$

lorsque $k \rightarrow \infty$.

PREUVE : En vertu d'un développement linéaire en série de Taylor,

$$\text{cov}\{k(\alpha_{1k} - \mu \alpha_{2k})^2, k^{1/2}(\alpha_{2k} - \alpha_2)\} = O(k^{-1/2})$$

Si les signes des α_k sont inversés pour $i = 1, 2$, alors $k(\alpha_{1k} - \mu \alpha_{2k})^2$ ne change pas, mais $k^{1/2}(\alpha_{2k} - \alpha_2)$ est inversé. Par conséquent, par symétrie,

$$k^{1/2}(\alpha_{2k} - \alpha_2) \bar{d} \stackrel{\text{P}}{=} N(0, \sigma_2^2) + k^{-1/2} \xi_k.$$

où χ_1^2 représente une variable chi carré aléatoire à un degré de liberté. En outre,

$$k(\alpha_{1k} - \mu \alpha_{2k})^2 \bar{d} \stackrel{\text{P}}{=} \sigma_2^2 \chi_1^2 + k^{-1/2} \xi_k.$$

pour certaines constantes σ_2^2 , alors

$$k^{1/2}(\alpha_{1k} - \mu \alpha_{2k}) \bar{d} \stackrel{\text{P}}{=} N(0, \sigma_2^2) + k^{-1/2} \xi_k$$

Maintenant, nous allons montrer que le terme de covariance de (A.1) est asymptotiquement négligeable. Puisque

$$\text{cov}\{(\alpha_{1k} - \mu \alpha_{2k})^2, (\alpha_{2k} - \alpha_2)\} + k^{-2} \xi_k. \quad (\text{A.1})$$

$$E\bar{Q}\mu_k = \alpha_2^{-2} \text{var}(\alpha_{1k} - \mu \alpha_{2k}) - 2\alpha_2^{-3}$$

ce qui implique

$$(\mu_k - \mu)^2 = \alpha_2^{-2} (\alpha_{1k} - \mu \alpha_{2k})^2 [1 - 2\alpha_2^{-1} (\alpha_{2k} - \alpha_2)] + k^{-2} \xi_k.$$

Par conséquent,

$$= \alpha_2^{-1} (\alpha_{1k} - \mu \alpha_{2k}) [1 - \alpha_2^{-1} (\alpha_{2k} - \alpha_2)] + k^{-3/2} \xi_k.$$

$$+ (\alpha_{2k} - \alpha_2) \mu (\alpha_2)^{-2} - (\alpha_{1k} - \alpha_1) (\alpha_{2k} - \alpha_2) (\alpha_2)^{-2}$$

$$= (\alpha_{1k} - \alpha_1) (\alpha_{2k} - \alpha_2) \mu (\alpha_2)^{-1}$$

$$+ k^{-3/2} \xi_k$$

$$\frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial \alpha_1^2} \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} + 2(\alpha_{1k} - \alpha_1) (\alpha_{2k} - \alpha_2) \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial \alpha_1 \partial \alpha_2} \left[\frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2} \right]$$

$$\left[\frac{1}{2} (\alpha_{1k} - \alpha_1)^2 \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial \alpha_1^2} + (\alpha_{2k} - \alpha_2)^2 \frac{\partial^2 f(\alpha_1, \alpha_2)}{\partial \alpha_2^2} \right]$$

$$= (\alpha_{1k} - \alpha_1) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_1} + (\alpha_{2k} - \alpha_2) \frac{\partial f(\alpha_1, \alpha_2)}{\partial \alpha_2}$$

$$\mu_k - \mu = \alpha_1 k / \alpha_2 k - \alpha_1 / \alpha_2$$

Tableau 5
Biais et écarts-types des estimations du revenu des ménages, stratification a posteriori selon le sexe, l'âge et la race

Estimateur		Non stratifié a posteriori		Stratifié a posteriori	
		\bar{y}_1/\bar{y}_2	\bar{y}_1/\bar{y}_2	\bar{y}_1/\bar{y}_2	\bar{y}_1/\bar{y}_2
Biais agrégé		1 412\$	640\$	1 173\$	757\$
Biais téléphone		1 414\$	640\$	1 177\$	747\$
Biais deuxième phase		-2\$	0\$	-4\$	10\$
Biais théorique		789\$	0\$	463\$	0\$
Écart-type simulé		1 534\$	1 518\$	1 445\$	1 435\$
Écart-type estimé		1 537\$	1 521\$	1 448\$	1 438\$
Écart-type théorique		1 535\$	1 518\$	1 440\$	1 430\$
Racine de l'erreur quadratique moyenne		2 085\$	1 647\$	1 861\$	1 622\$

La valeur réelle du revenu moyen des ménages est 40 187\$. Notons que \bar{y}_1/\bar{y}_2 et \bar{y}_1/\bar{y}_2 ne dépendent pas de la stratification a posteriori, si bien que les résultats obtenus ici sont les mêmes que ceux obtenus au tableau 4. La taille des échantillons sélectionnés est égale à 500 et le nombre de simulations est égal à 100 000.

* Cette valeur est fondée sur (3,7), tandis que la valeur fondée sur (3,8) est 1 421\$.

7. DISCUSSION

Nous proposons ici d'utiliser de grandes bases de données à grande diffusion (c'est-à-dire les PUMS) pour élaborer un modèle de la propension p_k d'un ménage à avoir le téléphone. Nous avons utilisé, pour la Virginie en 1990, un modèle GLIM avec une fonction de lien bilogarithmique et les variables indépendantes « nombre de personnes », « mode d'occupation du logement », « date d'entrée dans le logement », « nombre d'automobiles », « revenu du ménage », « langue » et « race ».

Nous avons proposé d'utiliser les coefficients de pondération en fonction de l'abonnement au téléphone p_k pour réduire le biais de non-couverture qui entache les estimateurs en cas d'enquêtes téléphoniques. Nous pouvons nous attendre à ce que ce biais se manifeste lorsque la variable étudiée est liée à la possession d'un téléphone. Les exemples que nous avons choisis sont tous des variables de ce type et, par conséquent, les améliorations observées lorsque l'on utilise des coefficients de pondération en fonction de l'abonnement au téléphone sont plus importantes que l'on ne s'y attendrait pour des variables pour lesquelles l'association à la possession d'un téléphone est faible. Les coefficients de pondération peuvent être combinés à la stratification a posteriori. Nous avons constaté que l'utilisation de ce genre de coefficient de pondération réduit fortement le biais qui entache l'estimateur non stratifié a posteriori ainsi que l'estimateur stratifié a posteriori, que l'on faille procéder à la stratification a posteriori, que la taille de l'échantillon soit suffisamment grande pour que

la probabilité d'être vide soit négligeable pour toute strate échantillons de taille égale à 500, le nombre de strates a posteriori était assez limité. Il est certain que, si l'on disposait d'un échantillon suffisamment grand pour pouvoir stratifier a posteriori en fonction des mêmes variables indépendantes que celles utilisées pour élaborer les p_k , l'utilisation de coefficients de pondération en fonction de l'abonnement au service téléphonique n'offrirait qu'une amélioration négligeable comparativement à la stratification a posteriori. Cependant, nombre de sondages d'opinion téléphoniques de portée nationale sont réalisés auprès d'un échantillon d'environ 1 000 personnes et nous sommes convaincus que, pour des échantillons de cette taille, l'application de coefficients de pondération en fonction de l'abonnement au téléphone permet de réduire considérablement le biais dû à l'échantillonnage téléphonique, à condition que les catégories soient corrigées comme il convient pour l'inflation.

Enfin, les PUMS sont ventilés par État et par grande région métropolitaine, de sorte que l'on peut créer des modèles pondérés en fonction de l'abonnement au téléphone distincts pour les grandes unités géographiques, mesure qui semble appropriée en cas de grande enquête.

REMERCIEMENTS

Les auteurs sont très reconnaissants à un évaluateur anonyme de ses nombreuses suggestions constructives.

ANNEXE : ÉTABLISSEMENT DES ÉQUATIONS

Avant de montrer comment ont été obtenues les équations présentées aux sections 3 et 4, nous devons supposer que les séries $\{\alpha_{11}, \alpha_{12}, \dots\}$, pour $i = 1, 2$ obéissent à certaines conditions de régularité. En outre, nous devons prouver certains lemmes. Ensuite, nous pourrions produire les équations des estimateurs \hat{p}_k et \hat{p}_k^{psw} . Chaque fois que nous introduisons la variable d'erreur ξ_k dans ce qui suit, alors $\xi_k = O_p(1)$ et $E(\xi_k^2) = O(1)$ quand $k \rightarrow \infty$. Afin de simplifier la notation

a postérieur et de 29 % celle de l'estimateur stratifié à postérieur. Si l'on prend l'EQM comme critère, l'amélioration due à la stratification à postérieur est moins importante.

Au tableau 5, nous estimons de nouveau le revenu du ménage, mais celle fois-ci, nous procédons à une stratification à postérieur selon le sexe (masculin, féminin), l'âge (moins de 45 ans, au moins 45 ans) et la race (noire, autre) du chef du ménage. Notons que cette nouvelle stratification à postérieur n'affecte pas les estimateurs non stratifiés à postérieur. Les estimations du *biais agrégé* fondées sur 100 000 simulations sont 1 173\$ et 757\$, et celles du *biais de téléphone* sont égales à 1 177\$ et 747\$ pour les estimateurs stratifiés à postérieur $\hat{\mu}^{psw}$ et $\hat{\mu}^{ps}$ respectivement. De nouveau, les valeurs du *biais de deuxième phase* sont faibles comparativement à celles du *biais de téléphone*. L'utilisation des p_k réduit de 35 % le biais du simplement à la stratification à postérieur.

Pour l'estimateur stratifié à postérieur, le biais théorique est égal à 463\$. Les valeurs de l'écart-type des simulations sont 1 445\$ et 1 435\$ pour les estimateurs $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ respectivement. Les racines carrées de l'erreur quadratique moyenne sont égales à 1 861\$ et 1 622\$ pour les estimateurs $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ respectivement. Donc, l'utilisation des p_k réduit de 24 % la valeur de l'EQM pour l'estimateur stratifié à postérieur.

La valeur de l'EQM de $\hat{\mu}^{psw}$ est à peu près la même au tableau 4 et au tableau 5. Cependant, l'EQM de $\hat{\mu}^{ps}$ diminue légèrement lorsqu'on passe du tableau 4 au tableau 5.

Tableau 4

Biais et écart-types des estimations du revenu des ménages, stratification à postérieur selon la race

Estimateur		Non stratifié		Stratifié à postérieur	
		$\hat{\mu}^{ps}$	$\hat{\mu}^{psw}$	$\hat{\mu}^{ps}$	$\hat{\mu}^{psw}$
Biais agrégé		1 412\$	640\$	1 192\$	633\$
Biais téléphone		1 414\$	640\$	1 193\$	630\$
Biais deuxième phase		-2\$	0\$	-2\$	3\$
Biais théorique		789\$	0\$	586\$	0\$
Ecart-type simulé		1 534\$	1 518\$	1 502\$	1 488\$
Ecart-type estimé		1 537\$	1 521\$	1 506\$	1 491\$
Ecart-type théorique		1 535\$	1 518\$	1 503\$	1 488\$
Racine de l'erreur quadratique moyenne		2 085\$	1 647\$	1 918\$	1 617\$

La valeur réelle du revenu moyen des ménages est 40 187\$. Notons que \hat{y}_1/\hat{y}_2 et $\hat{\mu}^{psw}$ ne dépendent pas de la stratification à postérieur, si bien que les résultats obtenus ici sont les mêmes que ceux obtenus au tableau 5. La taille des échantillons sélectionnés est égale à 500 et le nombre de simulations est égal à 100 000.

* Cette valeur est fondée sur (3.7), tandis que la valeur fondée sur (3.8) est 1 490\$.

qu'il est nécessaire d'éliminer la variable étudiée (ici le nombre d'automobiles) du modèle GLIM ajusté au PUMS pour que l'analyse soit appropriée.

Tableau 3

Biais et écarts-types des estimations du nombre moyen d'automobiles par ménage

Estimateur		Non stratifié à postérieur		Stratifié à postérieur	
		$\hat{\mu}^{ps}$	$\hat{\mu}^{psw}$	\hat{y}_1/\hat{y}_2	$\hat{\mu}^{psw}$
Biais agrégé		0,04872	0,01629	0,02226	0,01471
Biais téléphone		0,04872	0,01623	0,02220	0,01458
Biais deuxième phase		0,00000	0,00006	0,00089	0
Biais théorique		0,03388	0	0,00859	0
Ecart-type simulé		0,04694	0,04764	0,04162	0,04172
Ecart-type estimé		0,04682	0,04753	0,04148	0,04158
Ecart-type théorique		0,04715	0,04791	0,04152	0,04161
Racine de l'erreur quadratique moyenne		0,06765	0,05035	0,04720	0,04424

Le nombre moyen réel d'automobiles par ménage est égal à 1,8037. L'estimation à postérieur fondée sur le revenu. La taille des échantillons sélectionnés est de 500 et le nombre de simulations, de 100 000.

* Cette valeur du nombre moyen d'automobiles est fondée sur (3.7), tandis que la valeur basée sur (3.8) est égale à 0,04142.

6.3 Estimation du revenu moyen des ménages

Calculé pour le PUMS au 1/20 entier pour la Virginie en 1990, le revenu moyen des ménages est égal à 40 187\$. De nouveau, nous estimons les p_k mais cette fois-ci, c'est la covariable « revenu » que nous excluons du modèle GLIM ajusté au PUMS pour 1990, puisque c'est le revenu moyen des ménages que nous estimons.

Au tableau 4, pour estimer la valeur du revenu des ménages d'après un échantillon aléatoire simple de taille égale à 500 et d'après $\hat{\mu}^{ps}$ ou $\hat{\mu}^{psw}$, nous n'effectuons une stratification à postérieur que selon la race (noire, autre) du chef du ménage. Les estimations du *biais agrégé* basées sur 100 000 simulations sont 1 412\$, 640\$, 1 192\$ et 633\$, respectivement, et celles du *biais de téléphone* sont 1 414\$, 640\$, 1 193\$ et 630\$ pour les estimateurs \hat{y}_1/\hat{y}_2 , $\hat{\mu}^{ps}$, $\hat{\mu}^{psw}$ et $\hat{\mu}^{psw}$ respectivement. Donc, les valeurs du *biais de deuxième phase* sont faibles comparativement à celles du *biais de téléphone*. Dans l'ensemble, l'utilisation des p_k réduit de 55 % le biais produit par l'estimateur non stratifié à postérieur et de 47 % celui produit par l'estimateur stratifié à postérieur.

Les valeurs de l'écart-type des simulations sont 1 534\$, 1 518\$, 1 502\$ et 1 488\$, respectivement. Donc, les valeurs de la racine de l'erreur quadratique moyenne pour les quatre estimateurs sont environ égales à 2 085\$, 1 647\$, 1 918\$ et 1 617\$, respectivement, si bien que l'utilisation des p_k réduit de 38 % l'EQM de l'estimateur non stratifié

d'abonnés au téléphone, U^p pour calculer y_1/y_2 , $\hat{\mu}^w$, $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$. Ce biais est dû au fait que l'on échantillonne U^T plutôt que U . Dans la suite de l'exemple, nous adoptons la convention de mentionner entre parenthèses les estimations fondées sur les p_k de 1980 si ces estimations diffèrent de valeurs du *biais de deuxième phase* sont 0,01472, 0,00720 (0,00577), 0,01463 et 0,00850 (0,00838), et s'approchent donc de celles du *biais agrégué*.

Le *biais de deuxième phase*, qui correspond à la différence entre le *biais agrégué* et le *biais de téléphone*, est dû au fait que l'estimateur est l'approximation d'un ratio. Les estimations de ce *biais de deuxième phase*, tenant compte de l'erreur d'arrondissement, pour y_1/y_2 , $\hat{\mu}^w$, $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ sont 0,00000, 0,00002 (0,00001), -0,00002 et 0,00024 (0,00018), respectivement. Donc, dans notre exemple, le *biais de deuxième phase* est négligeable comparativement au *biais de téléphone*.

Les estimations du *biais théorique*, basées sur (3.2) de y_1/y_2 et $\hat{\mu}^{ps}$ sont 0,00777 (0,00905) et 0,00663 (0,00678), respectivement. Ces valeurs diffèrent de celles du *biais agrégué*, puisque (3.2) se base sur toutes les populations d'abonnés au téléphone possibles, tandis que le *biais agrégué* est subordonné à la réalisation en question de la population de basés sur le modèle voulant que chaque ménage soit caractérisé par une probabilité p_k de posséder le téléphone et sa valeur dépend donc du modèle utilisé pour l'ajustement au p_k . Puisque $\hat{\mu}^w$ et $\hat{\mu}^{psw}$ sont asymptotiquement non biaisés, leur *biais théorique* est défini comme étant nul.

Les valeurs de l'écart type simulé des 100 000 estimations simulées de μ pour y_1/y_2 , $\hat{\mu}^w$, $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ sont 0,01683, 0,01737 (0,01734), 0,01605 et 0,01643 (0,01634). Ces quatre nombres sont assez proches des valeurs de l'écart type estimé, qui sont égales à la racine carrée de la variance estimée moyenne de l'estimateur de μ , basées sur (3.1), (3.6) et (3.7). Précisément, ces valeurs de l'écart type estimé sont 0,01680, 0,01734 (0,01732), 0,01601 et 0,01635 (0,01628), respectivement. L'écart type estimé de rechange, fondé sur (3.8), de $\hat{\mu}^{psw}$ est 0,01610 (0,01606), valeur qui, de nouveau, s'approche de la valeur 0,01635 (0,01628). Les estimations de l'écart type *théorique* sont 0,01700 (0,01697), 0,01752 (0,01749), 0,01617 (0,01621) et 0,01658 (0,01653), si l'on se fonde sur les PUMS au 1/20 entier pour la Virginie pour 1990 et sur les équations (4.3), (4.4) et (4.5). Ces valeurs de l'écart type *théorique* s'approchent aussi des autres écarts types calculés.

L'utilisation des p_k réduit de 51 % (61 %) le *biais agrégué* de l'estimateur non stratifié a posteriori et de 40 % (41 %) celui de l'estimateur stratifié a posteriori. Toutefois, l'écart-type augmente légèrement. Les estimations de la racine carrée de l'erreur quadratique moyenne sont calculées d'après le *biais agrégué* et l'écart-type simulé des estimations y_1/y_2 , $\hat{\mu}^w$, $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ sont 0,02236 (0,02236),

6.2 Estimation du nombre moyen d'automobiles par ménage

Calculé d'après le PUMS au 1/20 entier pour la Virginie pour 1990, le nombre moyen d'automobiles par ménage est égal à 1,80397. La stratification a posteriori a été réalisée selon le revenu du ménage, en utilisant les catégories (moins de 20 000\$, au moins 20 000\$ mais moins de 35 000\$ et au moins 35 000\$). Nous excluons la covariable les p_k , mais celle fois-ci, nous excluons la covariable « nombre d'automobiles » du modèle GLIM ajusté au PUMS de 1990, puisque le nombre moyen d'automobiles est la variable que nous estimons.

Comme le montre le tableau 3, les estimations du *biais agrégué* sur 100 000 simulations d'échantillons aléatoires simples de taille égale à 500 sont 0,04872, 0,01629, 0,02226 et 0,01471, et celles du *biais de téléphone* sont 0,04872, 0,01623, 0,02220 et 0,01458 pour les estimateurs y_1/y_2 , $\hat{\mu}^w$, $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$, respectivement. Par conséquent, les valeurs du *biais de deuxième phase* sont assez faibles. L'utilisation des p_k réduit de 67 % le biais de l'estimateur non stratifié a posteriori et de 34 % celui produit par l'estimateur stratifié a posteriori. Il se pourrait que cette dernière valeur de la proportion de biais éliminée soit plus faible que la première parce que le revenu est un prédicteur puissant de l'abonnement ou non d'un ménage au service téléphonique (consulter Groves 1989, pages 116 à 119; Thornberry et Massay 1988), et que les groupes résultant de la stratification a posteriori pour la détermination de $\hat{\mu}^{ps}$ et $\hat{\mu}^{psw}$ sont fondés sur le revenu.

Les écarts-types des simulations sont 0,04694, 0,04764, 0,04162 et 0,04172, respectivement. Les valeurs de la racine de l'erreur quadratique moyenne pour les quatre estimateurs sont égales à environ 0,06765, 0,05035, 0,04720 et 0,04424, respectivement, si bien que l'utilisation des p_k réduit de 45 % l'EQM de l'estimateur non stratifié a posteriori et de 12 % celle de l'estimateur stratifié a posteriori.

Nous avons également réalisé des simulations, dont les résultats ne sont pas résumés dans les tableaux, où nous avons retenu la variable « nombre d'automobiles » lors de l'ajustement du modèle GLIM au PUMS de 1990. Les estimations du *biais agrégué* pour les estimateurs $\hat{\mu}^w$ et $\hat{\mu}^{psw}$ sont 0,00116 et 0,00006, respectivement, c'est-à-dire des valeurs nettement plus faibles que 0,01629 et 0,01471, qui sont les valeurs respectives du *biais relatif* obtenues en éliminant la variable « nombre d'automobiles » lors de l'ajustement du modèle GLIM. En outre, nous estimons

Méthode de rechange pour l'estimation des p_k

On pourrait poser aux participants à une enquête téléphonique par composition aléatoire les deux questions suivantes : « 1) Au cours des 12 derniers mois, combien de lignes téléphoniques votre ménage possédait-il? 2) Au cours des 12 derniers mois, pendant combien de mois chaque ligne téléphonique a-t-elle été en service? » Maintenant, représentons par p_k la somme des réponses à la question (2). Par exemple, dans un ménage possédant deux lignes téléphoniques, où l'une de ces lignes a été en service pendant les 12 mois et l'autre pendant 5 mois seulement, l'estimation de p_k serait de $12+5=17$. De nouveau, ici, p_k représente un coefficient de pondération plutôt qu'une probabilité. Poser cette deuxième question au répondant est une démarche comparable à celle recommandée par Brick et coll. (1994) qui ont également proposé de pondérer les données pour tenir compte de la probabilité qu'un ménage soit abonné au téléphone.

6. INFÉRENCES D'APRÈS LES VARIABLES DU MÉNAGE ET LES VARIABLES PERSONNELLES

Nous allons comparer les quatre estimateurs proposés de μ lorsque nous faisons des inférences concernant le taux de titulaires d'un diplôme d'études secondaires parmi les personnes de 21 ans et plus, le nombre moyen d'automobiles par ménage et le revenu moyen des ménages dans l'État de Virginie. Nous avons réalisé 100 000 simulations d'échantillons aléatoires simples comportant 500 ménages ayant le téléphone à partir du PUMS au 1/20 pour la Virginie pour 1990 en n'utilisant qu'une seule strate (c'est-à-dire $H=1$).

À la section 6.1, nous utilisons deux ensembles de valeurs de p_k . L'un se fonde sur l'ajustement d'un modèle de régression GLIM au PUMS de 1990 et l'autre sur l'ajustement d'un modèle GLIM au PUMS de 1980 en se servant, pour ajuster les catégories de revenu, du ratio du revenu médian du ménage de 1990 (32 800\$) au revenu médian du ménage de 1980 (17 510\$). L'utilisation de la valeur de p_k pour 1980 en vue d'estimer un paramètre pour 1990 montre à quel point notre méthode donne de bons résultats lorsque l'on utilise les coefficients du modèle GLIM pour les ensembles de données futures, à condition de faire à une correction pour tenir compte de l'inflation. Aux sections 6.2 et 6.3, nous utilisons uniquement les valeurs de p_k pour 1990.

Lorsque la taille des échantillons est suffisamment grande, on devrait recourir à la stratification a posteriori. Nous pouvons comparer entre-eux les estimateurs non stratifiés a posteriori, d'une part, et les estimateurs stratifiés a posteriori, d'autre part. La comparaison de y_1/y_2 à μ^{psw} est appropriée, ainsi que celle de μ^{ps} à μ^{psw} . Ces comparaisons montrent les améliorations dues à l'utilisation de p_k dans les estimateurs.

6.1 Estimation du taux de titulaires d'un diplôme d'études secondaires

Si l'on utilise l'entière de la population du PUMS au 1/20 pour la Virginie en 1990, le taux moyen de titulaires d'un diplôme d'études secondaires pour l'ensemble des Virginiens de 21 ans et plus est $\mu=0,75118$. Pour estimer le taux de titulaires d'un diplôme d'études secondaires au moyen d'un échantillon aléatoire simple et de μ^{ps} ou μ^{psw} nous procédons à une stratification a posteriori selon le sexe (masculin, féminin), l'âge (moins de 45 ans, 45 ans et plus) et la race (noire, autre) du chef du ménage. Nous estimons les valeurs de p_k d'après le tableau 1. Les valeurs des biais et des écarts-types dont nous disposons plus loin sont présentées au tableau 2 pour les p_k de 1990.

Tableau 2

Biais et écarts-types des estimations du taux de titulaires d'un diplôme d'études secondaires

Estimateur		Stratifié a posteriori	
		μ^{ps}	μ^{psw}
Biais agréégé	0,01471	0,00722	0,01461
Biais téléphone	0,01472	0,00720	0,01463
Biais deuxième phase	0,00000	-0,00002	0,00024
Biais théorique	0,00777	0	0,00663
Écart-type simulé	0,01683	0,01737	0,01605
Écart-type estimé	0,01680	0,01734	0,01601
Écart-type théorique	0,01700	0,01752	0,01617
Racine de l'erreur quadratique moyenne	0,02236	0,01881	0,02171

Le taux réel de titulaires d'un diplôme d'études secondaires est 0,75118. La stratification a posteriori est fondée sur le sexe, l'âge et la race. La taille des échantillons sélectionnés est de 500 et le nombre de simulations exécutées est de 100 000.

* Cette valeur est basée sur l'équation (3.7), tandis que la valeur obtenue en se basant sur (3.8) est 0,01610.

Pour estimer le biais agréégé de chacun des quatre estimateurs de μ , nous calculons la moyenne sur 100 000 simulations de la différence entre l'estimation pour un échantillon de taille 500 et μ . L'estimation du biais agréégé produit par y_1/y_2 , μ^{ps} et μ^{psw} sont 0,01471, 0,00722, 0,01461 et 0,00874, respectivement, si l'on utilise les valeurs de p_k pour 1990. Donc, l'utilisation des p_k réduit de 51 % le biais de l'estimateur non stratifié a posteriori et de 40 % celui de l'estimateur stratifié a posteriori.

L'utilisation des valeurs de p_k pour 1980 produit des résultats comparables. Les estimations du biais agréégé produit par μ^{ps} et μ^{psw} sont 0,00578 et 0,00856, respectivement, si l'on se sert des valeurs de p_k pour 1980. Cependant, ces résultats ne sont pas résumés dans les tableaux.

Le biais de téléphone présente dans le tableau 2 est le biais obtenu lorsque l'on échantillonne la population totale

Une correction recommandée si l'on recourt à la composition aléatoire consiste à demander à chaque répondant combien de lignes téléphoniques possède le ménage et à multiplier le nombre indiqué par l'estimation de p_k établie d'après le tableau 1 pour obtenir une nouvelle estimation de p_k . Par conséquent, p_k est maintenant un coefficient de pondération plutôt qu'une probabilité. Pour les simulations dont nous discutons à la section 6, cette correction n'est pas nécessaire, puisqu'il s'agit de ménages sélectionnés par échantillonnage aléatoire simple du PUMS, quel que soit le nombre de lignes téléphoniques. Donc, on peut se servir du tableau 1 pour estimer les p_k lorsque l'on réalise des enquêtes téléphoniques. Si l'on utilise un modèle de régression linéaire généralisé calculé d'après le PUMS produit pour une date de référence antérieure pour analyser les données d'une enquête ultérieure, on peut procéder à un rééchantillonnage afin de tenir compte des modifications de la distribution des revenus des ménages au cours du temps. Le tableau 1 donne aussi les coefficients d'un modèle calculé en se basant sur le PUMS de 1980. Nous donnons à la section 6 un exemple d'utilisation d'un modèle fondé sur un PUMS de 1980 pour calculer les p_k pour un échantillon tiré de la population du PUMS de 1990. Nous notons que le PUMS de 1980 ne contenait pas les covariables (date d'entrée dans le logement) et qu'on obtenait un modèle mieux ajusté lorsque les catégories de revenu utilisaient « Espagnol » et « Autre » combinées. En outre, le revenu médian du ménage a presque doublé de 1980 à 1990, si bien que le nombre de catégories de revenu utilisées en 1980 était plus faible qu'en 1990.

Bien que l'utilisation du tableau 1 soit pratique lorsque l'on procède à un échantillonnage à partir d'un PUMS et que l'on exécute des simulations, les données sur les covariables énumérées dans ce tableau pourraient ne pas être disponibles dans le cas d'enquêtes réelles où l'on recourt à la composition aléatoire. Il est possible de reproduire le tableau 1 en utilisant d'autres covariables ou d'estimer les p_k selon la méthode de rechange qui suit.

Si le nombre de personnes est supérieur à 5, le convertir à la valeur 5. La covariable « mode d'occupation du logement » ne figurait pas dans le PUMS de 1980. La catégorie « 40 000\$ à 49 999\$ » de 1980 englobe effectivement « 40 000\$ et plus ». La catégorie « autre » pour la

Tableau 1

Covariable	Catégorie	Valeur de 1990	Valeur de 1980
Nombre de personnes	0,2747	-0,0022	0,1929
Mode d'occupation du logement	Propriétaire	-0,5552	-0,7845
	Locataire	(0,0000)	(0,0000)
Date d'entrée dans le logement	1989 ou 1990	0,9742	NA
	1985 à 1988	0,5920	NA
	1980 à 1984	0,3489	NA
	1970 à 1979	0,2185	NA
	1969 ou avant	0,0000	NA
Nombre d'automobiles	0	1,2927	0,8633
	1	0,6842	0,3981
	2	0,1896	0,0399
	3 ou plus	0,0000	0,0000
Revenu	Moins de 0\$	3,5325	2,3639
	0\$ à 9 999\$	3,7929	2,5238
	10 000\$ à 19 999\$	3,4878	1,9763
	20 000\$ à 29 999\$	3,0299	1,0220
	30 000\$ à 39 999\$	2,4297	0,3889
	40 000\$ à 49 999\$	1,8899	0,0000
	50 000\$ à 59 999\$	1,5992	(0,0556)
	60 000\$ à 69 999\$	1,2144	(0,0543)
	70 000\$ à 79 999\$	1,0004	(0,0539)
	80 000\$ et plus	0,0000	(0,0538)
Langue	Anglais	0,6156	0,4232
	Espagnol	0,4889	NA
	Autre	0,0000	0,0000
Race	Noire	-0,4233	-0,3837
	Autre	0,0000	(0,0058)
Ordonnée à l'origine		-7,6707	-4,9024
		(0,0588)	(0,0322)

4.3 Estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone

La variance asymptotique théorique de l'estimateur stratifié a posteriori pondéré en fonction de l'abonnement au téléphone de μ est

$$\text{var } \hat{\mu}^{\text{psw}} = (N\alpha_2)^{-2} \sum_H \sum_{g=1}^H$$

$$\sum_{k \in \mathcal{K}_k} p_k \left[\frac{\left(\sum_{j \in \mathcal{K}_k} p_j \right) \left(\sum_{j \in \mathcal{L}_k^h} p_j \right) - n_h}{\left(\sum_{j \in \mathcal{L}_k^h} p_j \right) - 1} \right] \left(y_{1j} - \mu y_{2j} \right)^2$$

(4.5) $+ O(n^{-2} + N^{-1}).$

Puisque $\hat{\mu}^{\text{psw}}$ est asymptotiquement non biaisé, son EQM est identique au deuxième membre de l'équation (4.5).

5. ESTIMATION DES p_k AU MOYEN DES ÉCHANTILLONS DE MICRODONNÉES À GRANDE DIFFUSION

Le United States Bureau of the Census a produit les échantillons de microdonnées à grande diffusion (PUMS), c'est-à-dire des échantillons au 1/100 et au 1/20 de la population de chacun des 50 États et de Washington, D.C., pour l'année de référence 1990. Pour chaque personne sélectionnée dans l'échantillon, le fichier contient les données sur 75 variables du ménage et 75 variables personnelles, chaque ménage ayant un chef de ménage clairement défini. Nous utilisons les PUMS pour deux raisons. À la présente section, nous nous en servons pour estimer les p_k tandis qu'à la section 6, nous les utilisons pour l'exécution de simulations en vue de construire des exemples permettant de comparer les estimateurs.

Dans le présent article, nous utilisons l'échantillon au 1/20 pour la Virginie. Puisque 1/20, c'est-à-dire 5 %, représente un nombre énorme de ménages, nous traitons cet échantillon comme s'il s'agissait de la population totale de la Virginie. Comme nous sélectionnerons les ménages à partir de cet échantillon au 1/20. Nous pouvons faire des inférences concernant les variables personnelles, telles que le taux de titulaires d'un diplôme d'études secondaires et les variables du ménage, telles que le nombre d'automobiles ou le revenu d'un ménage. Les fichiers des PUMS contiennent, pour chaque ménage, des données indiquant si celui-ci est abonné ou non au téléphone. Nous avons éliminé de notre étude tous les ménages pour lesquels le fichier contenait, pour la situation d'abonnement au téléphone, la mention

« sans objet ». Ces ménages correspondaient soit à des logements vacants soit à des logements collectifs (institutionnels ou non). Pour 1990, le nombre de ménages retenus est de 110 744, dont 104 606 possédaient le téléphone, donc, la proportion de ménages retenus ayant le téléphone est de 94,5 %.

Pour estimer p_k , qui est la probabilité, ou la proportion, que le $k^{\text{ème}}$ ménage ait le téléphone, nous utilisons la régression linéaire généralisée avec un lien logarithmique sur l'échantillon au 1/20 pour la Virginie, ainsi que les coefficients de pondération attribués aux ménages dans le fichier PUMS. McCullagh et Nelder (1991, pages 107 à 110) recommandent d'utiliser un lien logarithmique lorsque la valeur des probabilités s'approche de l'unité et nous avons constaté que ce lien produit un bon ajustement du modèle. Nous avons également observé que la fonction de lien logit donne un mauvais ajustement.

Nous utilisons les coefficients de pondération des ménages du PUMS pour estimer les p_k , mais nous ne les utilisons nulle part ailleurs. En particulier, à la section 6, lors de la création d'échantillons de Monte Carlo de la population PUMS, il s'agit d'échantillons aléatoires simples de la population d'abonnés au téléphone.

Exemples d'estimation des p_k

Nous utilisons, pour estimer les p_k , six covariables, à savoir le nombre de personnes dans le ménage, le mode d'occupation du logement (propriétaire ou locataire), la date à laquelle le chef de ménage est entré dans le logement, le revenu du ménage, la langue parlée par le ménage et la race ainsi que les catégories pour chacune d'elles, en nous basant sur une analyse approfondie du PUMS de 1990 par les techniques de régression linéaire généralisée de SAS. L'analyse a montré que toutes ces covariables étaient hautement significatives. Pour obtenir les estimations des p_k , nous additionnons les estimations appropriées des covariables figurant dans le tableau 1. La covariable pour le nombre de personnes doit être multipliée par le nombre de personnes que compte le ménage; cependant, si le nombre de personnes est supérieur à 5, nous le fixons à 5 aux fins des calculs. Par exemple, si le ménage compte 3 Américains d'ascendance asiatique parlant l'anglais, possédant deux automobiles et vivant dans une maison achetée en 1987 et que le revenu du ménage est égal à 75 000\$, alors le tableau 1 indique que l'estimation de p_k est la solution de

$$\log(-\log p_k) = 3 \times 0,2747 - 0,5552 + 0,5920 \\ + 0,1896 + 1,0004 + 0,6156 + 0,0000.$$

Notons que, dans le tableau 1, pour les covariables de date d'entrée dans le logement, de nombre d'automobiles et de revenu, les valeurs correspondant aux catégories diminuent, comme prévu, de façon monotonique, sauf si le revenu est négatif.

Si la valeur de tout terme n_{gh} est faible, nous pourrions alors préférer l'estimateur

$$\widehat{\mu}^{\text{psw}} = [N \hat{a}^{\text{psw}(2)}]^{-2} \sum_H N^2 \left(\sum_{j \in U_h} p_j^{-1} \right)^{-2} \left(\frac{n_h - 1}{n_h} \right) \left(1 - \frac{n_h}{N} \right) \sum_G \sum_{k \in s_h} p_k^{-2} \left(\sum_{j \in s_h} p_j^{-1} \right)^{-1} \sum_{m=1}^{m \in s_h} p_m^{-1} (Y_{1m} - \hat{\mu}^{\text{psw}} Y_{2m})^2 \quad (3.8)$$

Notons que, si l'on connaissait la valeur des N_{Tgh} , situation fort peu probable, un estimateur mieux connu et plus intuitif de $\hat{\mu}^{\text{psw}}$ serait

$$\widehat{\mu}^{\text{psw}} = [N \hat{a}^{\text{psw}(2)}]^{-2} \sum_H \sum_{g=1}^G \frac{N_{Tgh}^2}{n_{gh}^2} \left(\sum_{k \in s_h} p_k^{-2} \right) \left(1 - \frac{n_{gh}}{N_{Tgh}} \right) \left(\sum_{j \in s_h} p_j^{-1} \right)^{-1} \sum_{m=1}^{m \in s_h} p_m^{-1} (Y_{1m} - \hat{\mu}^{\text{psw}} Y_{2m})^2 \quad (3.9)$$

Puisque ordinairement, on ne connaît pas les N_{Tgh} , (3.9) n'est généralement pas un estimateur pratique. Cependant, (3.9) facilite la justification de (3.7) et (3.8) qui sont, elles, des équations assez pratiques.

Puisque l'estimateur $\hat{\mu}^{\text{psw}}$ est asymptotiquement non biaisé, alors une estimation valide de l'EQM est identique à l'estimation de la variance. En outre, poser que $Y_{2j} = 1$ dans (3.7) et (3.8) permet d'estimer la variance de $\hat{a}^{\text{psw}(1)}$. Si $G = 1$, l'estimateur $\hat{\mu}^{\text{psw}}$ ne se réduit pas à $\hat{\mu}^w$, comme on pourrait naïvement s'y attendre. Lorsque $G = 1$ l'estimateur de choix est $\hat{\mu}^w$, puisqu'il est fondé sur un seul ratio, tandis que $\hat{\mu}^{\text{psw}}$ est fondé sur un ratio de ratios. L'utilisation de l'estimateur $\hat{\mu}^{\text{psw}}$ nécessite un échantillon de plus grande taille pour chaque catégorie strate-groupe, tandis que $\hat{\mu}^w$ nécessite uniquement un échantillon global de grande taille. Néanmoins, lorsque $H = G = 1$, les estimateurs $\hat{\mu}^w$ et $\hat{\mu}^{\text{psw}}$ sont identiques; les estimateurs de la variance fondés sur $\hat{\mu}^w$ sont préférables à ceux fondés sur $\hat{\mu}^{\text{psw}}$, car les estimations de la variance de $\hat{\mu}^{\text{psw}}$ sont basées sur un ratio de ratios.

4. ERREUR QUADRATIQUE MOYENNE ASYMPTOTIQUE

Nous allons maintenant donner l'erreur quadratique moyenne asymptotique des estimateurs de finis à la section 3. Les preuves, qui découlent de la linéarisation par série de Taylor, sont exposées à l'annexe, ainsi que les conditions mineures de régularité requises.

4.1 Estimateur stratifié à posteriori

Pour trouver la variance asymptotique théorique de l'estimateur stratifié à posteriori de μ , nous commençons par définir

$$a_i^* = \text{plim}_{n \rightarrow \infty} \hat{a}^{\text{ps}(i)} = N^{-1} \sum_H \sum_{g=1}^G N_{gh} \left(\sum_{j \in U_h} p_j^{-1} \right)^{-1} \sum_{k \in U_{gh}} p_k Y_{ik}, \quad (4.1)$$

pour $i = 1, 2$, ainsi que

$$\mu^* = a_1^* / a_2^* \quad (4.2)$$

Notons qu'en général, $a_i^* \neq a_i$ et $\mu^* \neq \mu$. La variance asymptotique théorique de $\hat{\mu}^{\text{ps}}$ est

$$\text{var } \hat{\mu}^{\text{ps}} = (N a_2^*)^{-2} \sum_H \sum_{g=1}^G$$

$$\frac{N_{gh}^2 \left(\sum_{j \in U_h} p_j^{-1} \right) \left(\sum_{k \in U_{gh}} p_k^{-1} \right) \left(\sum_{j \in U_h} p_j^{-1} \right)^{-1} \left(\sum_{k \in U_{gh}} p_k^{-1} \right)^{-1} \left(\sum_{j \in U_h} p_j^{-1} \right)^{-1} \left(\sum_{k \in U_{gh}} p_k^{-1} \right)^{-1}}{\sum_{k \in U_{gh}} p_k^{-2}}$$

$$\left[Y_{1k} - \mu^* Y_{2k} - \frac{\sum_{j \in U_h} p_j (Y_{1j} - \mu^* Y_{2j})}{\sum_{j \in U_h} p_j} \right]^2 + O(n^{-2} + N^{-1}) \quad (4.3)$$

lorsque $n \rightarrow \infty$. Nous avons montré dans (3.2) que le biais asymptotique de $\hat{\mu}^{\text{ps}}$ est $O(1)$ lorsque $n \rightarrow \infty$. Par conséquent, l'EQM asymptotique de $\hat{\mu}^{\text{ps}}$ est également $O(1)$ si $n \rightarrow \infty$.

4.2 Estimateur pondéré en fonction de l'abonnement au téléphone

La variance asymptotique théorique de l'estimateur pondéré en fonction de l'abonnement au téléphone de μ est

$$\text{var } \hat{\mu}^w = (N a_2^*)^{-2} \sum_H \sum_{g=1}^G \frac{\left(\sum_{j \in U_h} p_j^{-1} \right) \left(\sum_{k \in U_{gh}} p_k^{-1} \right) \left(\sum_{j \in U_h} p_j^{-1} \right)^{-1} \left(\sum_{k \in U_{gh}} p_k^{-1} \right)^{-1} \left(\sum_{j \in U_h} p_j^{-1} \right)^{-1} \left(\sum_{k \in U_{gh}} p_k^{-1} \right)^{-1}}{\sum_{k \in U_{gh}} p_k^{-2}}$$

$$\left[Y_{1k} - \mu Y_{2k} - \frac{\sum_{j \in U_h} p_j (Y_{1j} - \mu Y_{2j})}{\sum_{j \in U_h} p_j} \right]^2 + O(n^{-2} + N^{-1}). \quad (4.4)$$

Puisque $\hat{\mu}^w$ est asymptotiquement non biaisé, alors son EQM est identique au deuxième membre de l'équation (4.4).

qui est asymptotiquement non biaisé pour μ , puisque $\hat{\alpha}^{(i)}$ est un estimateur non biaisé de $\alpha^{(i)}$, pour $i = 1, 2$. Nous montrons qu'une estimation valide de la variance de $\hat{\mu}^w$ est

$$\widehat{\text{var}} \hat{\mu}^w = [N \hat{\alpha}^{(2)}]^{-2} \sum_{h=1}^H \frac{N_h^2}{N^2} [1 - (n_h^h / N^{Th})] \quad \text{donnée par}$$

$$\sum_{k \in s_h} \left[\frac{y_{1k} - \hat{\mu}^w y_{2k}}{y_{1k} - \hat{\mu}^w y_{2k}} - n_h^h \sum_{j \in s_h} \frac{p_j}{d_j} \right] \quad (3.6)$$

Puisque l'estimateur $\hat{\mu}^w$ est asymptotiquement non biaisé, alors une estimation valide de l'EQU de $\hat{\mu}^w$ est identique à l'estimation de la variance. Poser $y_{2j} = 1$ dans (3.4) et (3.5) permet qu'une estimation valide de $\alpha^{w(1)}$ soit

$$\hat{\alpha}^{w(1)} = \sum_H \frac{1}{N^{Th} n_h^h} \sum_{k \in s_h} p_k^{-1} \left[\sum_{h=1}^H \sum_{k \in s_h} N^{Th} n_h^h \sum_{k \in s_h} p_k^{-1} \sum_{k \in s_h} y_{1k} \right]$$

On peut estimer la variance de $\hat{\alpha}^{w(1)}$ en posant $y_{2j} = 1$ dans (3.6).

3.3 Estimateur stratifié à postériori pondéré en fonction de l'abonnement au téléphone

L'autre estimateur que nous proposons, qui combine la stratification à postériori et l'estimateur pondéré en fonction de l'abonnement au téléphone, est peut-être le meilleur des quatre si la taille des échantillons est suffisante pour justifier la stratification à postériori. Toutefois, pour utiliser ce nouvel estimateur, il faut que tous les N_h^{gh} soient suffisamment grands pour que, lorsque la probabilité est élevée, les n_h^{gh} ne soient pas trop petits. Pour estimer α_i , nous utilisons

$$\hat{\alpha}^{ps(i)} = N^{-1} \sum_H \sum_G \sum_{h=1}^G \sum_{k \in s_h} N^{gh} \left[\sum_{j \in s_h} p_j^{-1} \sum_{k \in s_h} p_k^{-1} y_{jk} \right]$$

pour $i = 1, 2$. Puis, nous estimons μ par $\hat{\mu}^{psw} = \hat{\alpha}^{ps(1)} / \hat{\alpha}^{ps(2)}$. L'estimation de la variance de $\hat{\mu}^{psw}$ est donnée par

$$\widehat{\text{var}} \hat{\mu}^{psw} = [N \hat{\alpha}^{ps(2)}]^{-2} \sum_H \sum_G \sum_{h=1}^G \sum_{k \in s_h} N^{gh} \left(\sum_{j \in s_h} p_j^{-1} \sum_{k \in s_h} p_k^{-1} \right)$$

deviennent $\hat{\alpha}^{ps(i)}$ et α_i , respectivement. Nous pouvons alors appliquer les équations (3.1) et (3.2) de sorte que

$$\widehat{\text{var}} \hat{\alpha}^{ps(1)} = N^{-2} \sum_H \sum_G \sum_{h=1}^G \sum_{k \in s_h} N^{gh} \left[\frac{n_h^{gh} (n_h^{gh} - 1)}{1 - (n_h^{gh} / N^{Th})} \right] \quad (3.3)$$

$$\sum_{k \in s_h} \left[y_{1k} - n_h^{gh} \sum_{j \in s_h} y_{1j} \right] \quad \text{et}$$

biases $\hat{\alpha}^{ps(1)}$

$$= N^{-1} \sum_H \sum_G \sum_{h=1}^G \sum_{k \in s_h} N^{gh} \left(\sum_{j \in s_h} p_j^{-1} \sum_{k \in s_h} p_k^{-1} y_{1k} - \alpha_1 + O(n^{-1}) \right)$$

$= O(1)$

lorsque $n \rightarrow \infty$. Cochran (1977, pages 134-135) a produit un facteur de correction, d'ordre n^{-2} , pour (3.3). Toutefois, ce facteur de correction ne s'applique pas à (3.1) puisque le terme d'erreur due à l'estimation d'après le ratio est

$O(n^{-2})$.

Comme d'habitude, l'estimateur par ratio, représenté par \hat{y}_1 / \hat{y}_2 , est défini comme étant le ratio de la moyenne d'échantillon de y_1 à la moyenne d'échantillon y_2 , c'est-à-dire

$$\hat{y}_1 / \hat{y}_2 = \sum_{k \in s} y_{1k} / \sum_{k \in s} y_{2k}$$

L'estimateur stratifié à postériori et l'estimateur par ratio sont identiques si $G = H = 1$. Puisque nous n'utiliserons qu'une seule strate (c'est-à-dire $H = 1$) à la section 6, il n'est pas nécessaire que nous fassions référence à une théorie distincte pour l'estimateur par ratio.

3.2 Estimateur pondéré en fonction de l'abonnement au téléphone

Puisque l'estimateur stratifié à postériori, $\hat{\mu}^{ps}$, est biaisé, nous proposons deux autres estimateurs. L'un est l'estimateur pondéré en fonction de l'abonnement au téléphone, qui tient compte de la probabilité qu'un individu ait le téléphone. À la présente section, nous supposons que les p_k sont connues pour toutes les valeurs de $k \in s$ ou que l'on peut les estimer exactement. Nous décrivons l'estimation des p_k d'après les PUMS à la section 5. Pour obtenir une estimation grossière de α_i pour $i = 1, 2$, utilisons

$$\hat{\alpha}^{w(i)} = N^{-1} \sum_H \sum_{h=1}^H N^{Th} n_h^h \sum_{k \in s_h} p_k^{-1} y_{ik} \quad (3.4)$$

Puis, estimons μ au moyen de

$$\hat{\mu}^w = \hat{\alpha}^{w(1)} / \hat{\alpha}^{w(2)} \quad (3.5)$$

déterminons des estimations raisonnables de ces variances. Puisqu'une moyenne de population est un cas spécial de ratio et qu'un total de population est un multiple de ratio, les résultats concernant les estimateurs des moyennes ou des totaux découlent des résultats concernant les estimateurs des ratios.

2. NOTATION

Considérons N ménages dans une population U . Pour chaque ménage compris dans U , représentons deux variables étudiées par y_{1k} et y_{2k} , pour $k \in U$. À tous points dans le temps, le fait que le $k^{\text{ème}}$ ménage possède ou ne possède pas le téléphone est traité comme étant aléatoire, tandis que y_{1k} est traité comme étant fixe.

En posant que

$$\alpha_i = N^{-1} \sum_{k \in U} y_{ik},$$

pour $i = 1, 2$, l'objectif est d'estimer α_1 , α_2 et le ratio

$$\mu = \alpha_1 / \alpha_2.$$

En toute généralité, nous nous concentrons sur l'estimation de α_1 et μ .

Un cas spécial important de l'estimation d'un ratio μ se pose lorsque l'on souhaite estimer la moyenne d'une variable z_k pour $k \in D$ pour une sous-population $D \subset U$ que l'on ne peut pas échantillonner directement. Les sous-populations définies selon la race en sont des exemples. Posons que x_k est égal à 1 si $k \in D$ et égal à 0 autrement. Posons aussi que $y_{1k} = z_k x_k$ et $y_{2k} = x_k$. Alors, μ représente la moyenne de population de z_k sur la sous-population D .

Supposons qu'il existe H strates et utilisons la notation g comme indice de la strate. Supposons qu'il existe des groupes et choisissons la notation g comme indice des groupes qui sont utilisés pour construire les strates à posteriori. Nous connaissons les strates avant l'échantillonnage, mais nous n'observons les groupes qu'après la sélection de l'échantillon final. Par conséquent, U^{gh} représente tous les ménages dans le groupe g et la strate h , N^{gh} représente la taille de U^{gh} et N_h représente la taille de U^h . Les autres variables sont définies de la même façon en fonction de g et de h .

Représentons par U_T la population de ménages dans U qui sont abonnés au téléphone à l'heure actuelle et représentons par N_T la taille de U_T . La probabilité, ou propension, que le $k^{\text{ème}}$ ménage dans U soit également dans U_T est représentée par p_k et nous supposons que $p_k > 0$ pour toute valeur de k . Un échantillon aléatoire simple de taille n_h est tiré à partir de U^{gh} pour $h = 1, \dots, H$. Représentons par s_h l'échantillon final dans la strate h . La taille de l'échantillon final, s , est représentée par n . Pour l'analyse asymptotique, ici, nous supposons que $n/N \rightarrow 0$ lorsque $n \rightarrow \infty$ dans le même esprit que Särndal, Swensson et Wretman (1992, pages 166 à 169).

3. LES ESTIMATEURS

Le plan d'échantillonnage est traité comme un plan à deux phases avec échantillonnage de Poisson à la première phase et échantillonnage aléatoire simple stratifié à la deuxième phase. Chaque individu entre dans la population d'abonnés au téléphone avec la probabilité p_k pour $k \in U$, puis entre dans l'échantillon final conformément à la sélection d'un échantillon aléatoire simple de taille n_h , $h = 1, \dots, H$. Nous supposons que les p_k sont connues ou qu'elles peuvent être estimées exactement, comme nous le montrons à la section 5. La validation des estimateurs de μ examinés à la présente section figure à l'annexe.

3.1 Estimateur stratifié à posteriori et estimateur par ratio

Les estimations stratifiées à posteriori de α_1 et α_2 sont

$$\hat{\alpha}^{ps(i)} = N^{-1} \sum_G \sum_{h=1}^H \sum_{k \in s_h} N^{gh} n^{gh} y_{ik},$$

pour $i = 1, 2$, et l'estimation stratifiée à posteriori de μ est $\hat{\mu}^{ps} = \hat{\alpha}^{ps(1)} / \hat{\alpha}^{ps(2)}$. On sait qu'une estimation valide de la variance, subordonnée à U_T , (consulter Särndal et coll. 1992, pages 270-271) est

$$\widehat{\text{var}}(\hat{\mu}^{ps} | U_T) = (N \hat{\alpha}^{ps(2)})^{-2} \sum_H \sum_{h=1}^H \sum_G \left[1 - (n^{gh} / N^{gh}) \right] \left[\frac{n^{gh} (n^{gh})}{N^{gh}} \right]$$

$$\sum_{k \in s_h} \left[y_{1k} - \hat{\mu}^{ps} y_{2k} - n^{gh} \sum_{j \in s_h} (y_{1j} - \hat{\mu}^{ps} y_{2j}) \right]^2. \quad (3.1)$$

Bien que l'on ne puisse estimer le biais d'après l'échantillon final, il est bien connu que le biais théorique de $\hat{\mu}^{ps}$ est

$$\text{biais } \hat{\mu}^{ps} = \frac{\sum_H \sum_{h=1}^H \sum_G N^{gh} \sum_{j \in U^{gh}} p_j \left[\sum_{k \in U^{gh}} p_k y_{2k} \right]^{-1} \sum_{k \in U^{gh}} p_k y_{1k}}{\sum_{k \in U} p_k y_{1k} - \sum_{k \in U} p_k y_{2k}} + O(n^{-1})$$

(3.2)

lorsque $n \rightarrow \infty$. Notons, en examinant (3.2), qu'en général, l'EQM de $\hat{\mu}^{ps}$ ne tend pas vers zéro à mesure que la taille n de l'échantillon augmente.

Pour déterminer la variance et le biais de $\hat{\alpha}^{ps(i)}$, posons $y_{2k} = 1$ pour toutes les valeurs de k , de sorte que $\hat{\mu}^{ps}$ et μ

téléphone que nous proposons. L'erreur quadratique moyenne (EQM) de l'estimateur pondéré en fonction de l'abonnement au téléphone et celle de l'estimateur stratifié à posteriori pondéré en fonction de l'abonnement au téléphone tendent vers zéro lorsque la taille de l'échantillon devient importante, contrairement à l'EQM des deux autres estimateurs.

Nous adoptons un modèle à deux phases pour nos quatre estimateurs. La première phase comprend la sélection de ménages à partir de la population entière pour faire partie de la population d'abonnés au téléphone. Nous traitons la propension d'un ménage à être abonné au service téléphonique comme étant la probabilité que le ménage soit sélectionné pour faire partie de la population ayant le téléphone et nous supposons que cette probabilité est positive (quoique, éventuellement, faible) pour chaque ménage. La deuxième phase est la sélection d'un échantillon aléatoire simple stratifié (peut-être selon des variables géographiques) de la population ayant le téléphone. Dans les exemples donnés à la section 6, nous considérons la stratification à posteriori selon des caractéristiques telles que la race et l'âge du chef du ménage. Puisque la taille de nos échantillons est faible, nous ne stratifions pas la population de la Virginie selon des variables géographiques, bien que notre formule permette à la fois la stratification et la stratification à posteriori.

Idealement, on procéderait à la stratification à posteriori en se servant des mêmes covariables que celles utilisées pour estimer la propension à être abonné au téléphone à la première phase du modèle. Dans ce cas, les trois estimateurs fondés sur la propension à être abonné au service téléphonique et (ou) la stratification à posteriori seront presque identiques. Cependant, la taille d'échantillon de chaque catégorie stratifiée à posteriori ne doit pas être faible, de sorte que des limites pratiques limitent le nombre de catégories qu'il convient d'utiliser pour la stratification à posteriori. Néanmoins, de nombreuses catégories peuvent être utilisées pour déterminer la propension à être abonné au service téléphonique d'après les PUMS, parce que l'on utilise l'entièreté de la population.

Même s'il est possible de procéder à une stratification à posteriori au moyen d'un grand nombre de covariables, les formules habituelles de la variance en cas de stratification à posteriori exigent que l'on tire un échantillon aléatoire stratifié à partir de la population entière. Or, dans notre situation, nous sélectionnons un échantillon aléatoire stratifié à partir de la population ayant le téléphone, si bien que les formules habituelles de la variance ne sont pas applicables. Les techniques mises au point par Politz et Simmons (consulter Cochran 1977, pages 374 à 377) demandent qu'on utilise la population entière comme base de sondage plutôt que simplement la population ayant le téléphone et ne s'appliquent donc pas à notre scénario puisque'il permet la non-couverture.

Nous calculons la variance asymptotique de chacun des quatre estimateurs d'un ratio de population et nous

Microdata Samples) du Recensement de 1990, échantillon qui représente 5 % de la population. Une variable du fichier indique si les ménages sont ou non abonnés au téléphone. Nous estimons la propension (probabilité) à être abonné au service téléphonique par régression linéaire généralisée avec fonction de lien bilogarithmique, car le lien logit produit un ajustement médiocre du modèle. Nous recommandons d'utiliser notre modèle ajusté de régression, ainsi que les paramètres estimés, pour estimer ces probabilités de façon générale, lorsqu'un échantillon aléatoire est sélectionné à partir de la population d'abonnés au téléphone de Virginie.

Parce qu'il s'agit d'énormes ensembles de données, les PUMS jouent un autre rôle utile dans la présente étude. Nous les utilisons pour comparer le biais et la variance des estimateurs en examinant l'ensemble de la population d'abonnés au téléphone et en tirant des échantillons réétés de cette population. Les PUMS contiennent des données nominales correspondant à 75 variables de ménages et 75 variables personnelles pour tous les membres des ménages sélectionnés.

Dans les exemples donnés à la section 6, nous estimons le taux de titulaires d'un diplôme d'études secondaires, le nombre moyen d'automobiles par ménage et le revenu moyen du ménage au moyen d'estimateurs stratifiés et non stratifiés à posteriori pour des échantillons de 500 unités provenant des PUMS. Pour le taux de titulaires d'un diplôme d'études secondaires, les variables de stratification à posteriori sont le sexe, l'âge et la race du chef du ménage. Pour le nombre moyen d'automobiles par ménage, les variables de stratification à posteriori sont le revenu du ménage uniquement. Nous analysons les estimateurs du revenu moyen du ménage deux fois, d'abord en utilisant comme variable de stratification à posteriori uniquement la race du chef du ménage, puis en utilisant comme variables de stratification à posteriori le sexe, l'âge et la race du chef du ménage. Chaque une de ces variables de stratification à posteriori est répartie en deux catégories, sauf le revenu, qui est réparti en trois catégories.

Le fait que les estimateurs ne tiennent pas compte de la propension à être abonné au téléphone pour grand inconvénient que ces estimateurs ne sont pas asymptotiquement non-biaisés lorsque la taille de l'échantillon devient importante. L'un des principaux objectifs du présent article est de montrer que le biais est réduit considérablement lorsque les estimateurs tiennent compte de la propension à être abonné au téléphone, telle qu'estimée d'après les PUMS. Puisque nous considérons les estimateurs stratifiés à posteriori ainsi que non stratifiés à posteriori, de même que l'usage ou le non-usage de la propension à être abonné et à pas être abonné au service téléphonique, nous examinons ici quatre estimateurs d'une moyenne de population qui sont la moyenne d'échantillon, l'estimateur stratifié à posteriori habituel, un estimateur pondéré en fonction de l'abonnement au téléphone et, enfin, l'estimateur stratifié à posteriori pondéré en fonction de l'abonnement au

Estimation améliorée des ratios dans le cas des enquêtes téléphoniques avec correction pour la non-couverture

STEVEN T. GARREN et TED C. CHANG¹

RÉSUMÉ

Comme certains membres d'une population peuvent ne pas avoir le téléphone, les enquêtes téléphoniques par la méthode de composition aléatoire à l'intérieur des strates donnent parfois lieu à des estimateurs asymptotiquement biaisés des ratios. Nous examinons l'effet associé à l'impossibilité d'échantillonner la population non abonnée au téléphone. Nous tenons compte de la proportion d'un ménage à avoir le téléphone pour proposer un estimateur stratifié à posteriori pondéré en fonction de l'abonnement au téléphone qui semble donner de meilleurs résultats que l'estimateur stratifié à posteriori type si l'on prend pour critère l'erreur quadratique moyenne. Nous estimons les proportions à posséder le téléphone d'après les échantillons de microdonnées à grande diffusion (PUMS pour *Public Use Microdata Samples*) provenant du Recensement des États-Unis. Nous considérons des estimateurs non stratifiés à posteriori lorsque la taille de l'échantillon est faible. Nous calculons, pour chaque échantillon, l'erreur quadratique moyenne asymptotique, ainsi que son estimation fondée sur un échantillon. Enfin, nous examinons des exemples réels en nous servant des PUMS. Les autres formes de non-réponse ne sont pas examinées dans le cadre de cette étude.

MOTS CLÉS : Asymptotique; échantillons de microdonnées à grande diffusion du Recensement; stratification à posteriori; enquête téléphonique.

1. INTRODUCTION

Considérons les enquêtes où l'on échantillonne la population d'abonnés au téléphone. Les problèmes importants que posent les enquêtes téléphoniques sont la non-réponse (c'est-à-dire le refus de participer à l'enquête) et la non-couverture (c'est-à-dire le non-abonnement au service téléphonique). La non-réponse peut être la cause d'un biais plus important que la non-couverture, puisque la proportion à ne pas répondre est habituellement nettement plus élevée que la proportion à ne pas être abonné au téléphone. Cependant, ici, nous n'examinons que brièvement le cas de la non-réponse, car l'étude se concentre sur la non-couverture.

1.1 Revue de la littérature

Khuri et al (1995) ont produit une bibliographie de grande envergure des articles traitant des enquêtes téléphoniques. Steeh, Groves, Comment et Hansman (1983, pages 189 à 197) donnent des exemples de taux de non-réponse. Little (1986) et Rubin (1987) examinent des méthodes de correction de la non-réponse, par pondération ou par imputation. Rao (1997) donne une vue d'ensemble des enquêtes par sondage, y compris une discussion des méthodes de rééchantillonnage, en particulier le jackknife pour l'estimation de variance. La discussion inclue des techniques pour l'estimation de la variance due à l'imputation. En ce qui concerne la non-couverture, Brick, Waksberg et Keeter (1994) constatent qu'aux États-Unis, en tout

¹ Steven T. Garren, Department of Mathematics and Statistics, MSC 7803, James Madison University, Harrisonburg, Virginia, 22807, U.S.A. Travaux de recherche financés en partie par la subvention N-00014-92-J-1009 de l'ONR. Charlottesville, Virginia, 22903, U.S.A. Travaux de recherche financés en partie par la subvention MH53259-01A2 du NIMH; Ted C. Chang, Division of Statistics, 108 Halsey Hall, University of Virginia.

À partir de plusieurs caractéristiques, telles que la propriété du logement et la langue parlée par les membres du ménage, nous estimons ici la proportion d'un ménage à être abonné au service téléphonique en nous servant de la partie correspondante à la Virginie des échantillons de microdonnées à grande diffusion (PUMS pour *Public Use* (1988) examinent la non-couverture pour divers groupes sociodémographiques de 1963 à 1986 et constatent que le revenu est le déterminant le plus important de la probabilité qu'un ménage ait le téléphone.

1.2 Notre approche

populations d'abonnés irréguliers et de ménages non-perpétuellement sans le téléphone. Ces similitudes entre les irréguliers » étaient comparables à celles des ménages (sauf la propriété du logement) des ménages « abonnés différentes périodes durant l'année en question. Il observe aussi que la plupart des caractéristiques socioéconomiques qui avaient été abonnés et non-abonnés au téléphone à la étaient des abonnés irréguliers, c'est-à-dire des ménages façon continue au service téléphonique durant cette année-plus de la moitié des ménages n'ayant pas été rattachés de mentionne que, durant une enquête réalisée de 1992 à 1993, est interrompu sont habituellement indigents. Keeter (1995) notent aussi que les ménages dont le service téléphonique temps, 94 % des ménages sont abonnés au téléphone. Ils

7. CONCLUSIONS

La technique que nous avons mise au point semble

permettre d'etalonner les paramètres du modèle logit et du

modèle de survie de façon à ce que les caractéristiques

essentielles des données de l'EPA soient reflétées par les

prévisions du modèle LifePaths. L'élément clé de l'étalon-

nage du modèle logit est l'ajustement du paramètre corres-

pondant aux « termes constants » dans le prédicteur linéaire

qui est intégré dans la fonction de distribution logistisque

afin de prédire l'espérance conditionnelle de la variable

dépendante. À la section 3.1, nous élaborons la technique

dans un cadre général englobant d'autres modèles de choix

binaires. Cette technique pourrait notamment être étendue

au modèle probit bien connu où un prédicteur linéaire est

intégré dans la fonction de distribution normale type. L'éta-

lonnage du modèle semiparamétrique de survie repose sur

l'ajustement de tous les paramètres représentant la fonction

de hasard de base. Nos résultats illustrent comment il est

possible de faire évoluer en fonction du temps la forme

entière de la distribution des durées prévues par le modèle

selon un profil révèle par des données supplémentaires.

REMERCIEMENTS

Les auteurs remercient Steve Gribble et les membres du

Groupe de la modélisation socioéconomique de Statistique

Canada pour leurs commentaires utiles durant le dévelop-

pement du module sur le congé de maternité. Geoff Rowe

et Huan Nguyen pour l'utilisation de leur programme

informatique afin de suivre les individus durant leur parti-

cipation à l'échantillon de l'EPA, Katherine Marshall pour

le partage de ses programmes informatiques. Adhrien ten

Cate pour les discussions fructueuses qu'ils ont eues avec

elle et un évaluateur anonyme pour plusieurs améliorations

proposées. Ces travaux ont été réalisés à l'époque où les

deux auteurs travaillaient pour le Groupe de la modélisation

socioéconomique, Statistique Canada, immeuble

BIBLIOGRAPHIE

APPLEBY, J., BOOTHBY, D., ROULEAU, M. et ROWE, G.

(1999). Level and Distribution of Individual Returns to

Post-Secondary Education: Simulation Results from the LifePaths

Model. Présenté à la réunion de Canadian Economics

Association.

CHAMBLESS, L.E., et BOYLE, K.E. (1985). Maximum Likelihood

methods for complex sample data: Logistic regression and discrete

Theory and Methods, 14, 177-192.

CHEN, E.J., et ODERKIRK, J. (1997). Varied Pathways: The

Undergraduate Experience in Ontario. Feature article. *Education*

Quarterly Review, Statistique Canada, 4, 3, 47-62.

- CHEN, E.J., et ROWE, G. (1999). Trend Correlation of Labour
Market Earnings in Canada: 1982 to 1995. *Statistical Society of
Canada 1999 Proceedings of the Survey Methods Section*,
173-179.
- COX, D.R., et OAKS, D. (1984). *Analysis of Survival Data*. New
York: Chapman and Hall.
- DENNIS, J.E. Jr, et SCHNABEL, R.B. (1983). *Numerical Methods
for Unconstrained Optimization and Nonlinear Equations*.
Englewood Cliffs, NJ: Prentice-Hall.
- HAN, A., et HANUSMAN, J. A. (1986). Semiparametric Estimation of
Duration and Competing Risk Models. M.I.T. Document de
travail, numéro 450.
- KAPLAN, E.T., et MEIER, P. (1958). Nonparametric estimation
from incomplete observations. *Journal of the American Statistical
Association*, 53, 457-81.
- MARSHALL, K. (1999). Employment after childbirth. *Perspectives
on labour and income*. Statistique Canada, 18-25.
- MEYER, B.D. (1990). Unemployment Insurance and Unemployment
Spells. *Econometrica*, 58, 757-782.
- MORIL, J.G. (1989). Regression logistisque selon des plans de
sondage complexes. *Techniques d'enquête* 15, 213-233.
- PLAGER, L., et CHEN, E.J. (1999). Student Debt from 1990-91 to
1995-96: An Analysis of Canada Student Loans Data. MAJOR
RELEASES, *THE DAILY and Education Quarterly Review*,
Statistique Canada, 5, 4, 10-35.
- PRENTICE, R., et GLOECKLER, L. (1978). Regression Analysis of
Grouped Survival Data with Application to Breast Cancer Data,
Biometrics, 34, 57-67.
- ROBERTS, G.A., RAO, J.N.K., et KUMAR, S. (1987). Logistic
regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROWE, G., et CHEN, E.J. (1998). Un modèle à entrées et sorties de
la progression scolaire au secondaire dans les provinces
Canadiennes. *Recueil : Symposium 98, Analyse longitudinale
pour les enquêtes complexes*, Statistique Canada, 183-192.
- ROWE, G., et LIN, X. (1999). Modélisation des carrières au sein de
la population active pour le modèle de simulation des *LifePaths*,
*Recueil : Symposium 99, Combiner des données de sources
différentes*, Statistique Canada, 61-69.
- WOLFFSON, M.C. (1997). *Sketching LifePaths: A New Framework
for Socio-Economic Statistics. Stimulating Social Phenomena*,
(Eds. Conte, R. Geselemann et P. Terna). Lecture Notes in
Economics and Mathematical Systems, 456, Springer.
- WOLFFSON, M.C., et ROWE, G. (1996). Perspectives on Working
Time Over the Life Cycle, Canadian Employment Research Forum
Conference on Changes to Working Time, Ottawa.
- WOLFFSON, M.C., et ROWE, G. (1998a). LifePaths - Toward an
Integrated Microanalytic Framework for Socio-Economic
Statistics. 26th General Conference of the International
Association for Research in Income and Wealth, Cambridge, U.K.
- WOLFFSON, M.C., et ROWE, G. (1998b). Public Pension Reforms
- Analyses Based on the LifePaths General Accounting
Framework, 26th General Conference of the International
Association for Research in Income and Wealth, Cambridge, U.K.
- WOLFFSON, M.C., ROWE, G., GRIBBLE, S. et LIN, X. (1998).
Historical General Accounting with Heterogeneous
Populations. *Government Finance and Generational Equity* (Ed.
M. Corak), Statistique Canada, 107-127.

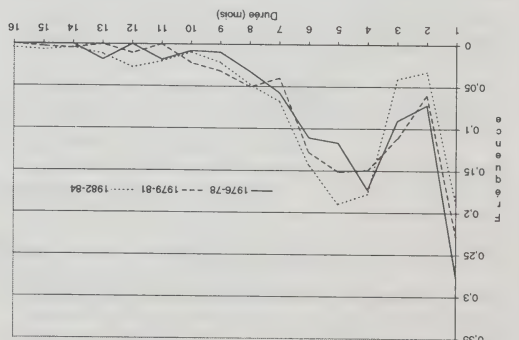


Figure 3. Données de l'EPA : Distribution des durées des congés pour 1976 à 1984

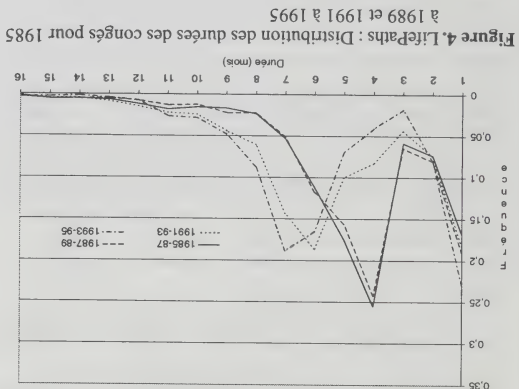
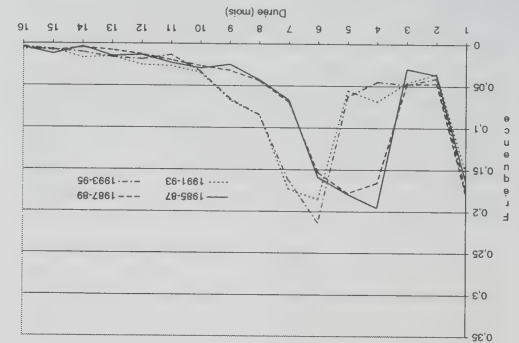


Figure 4. LifePaths : Distribution des durées des congés pour 1985 à 1989 et 1991 à 1995

Figure 5. Données de l'EPA : Distribution des durées des congés pour 1985 à 1989 et 1991 à 1995



À la figure 7, nous présentons la durée moyenne des congés de maternité débutant chaque année de la période observée. Nous comparons la moyenne des durées simulées à celles des durées calculées d'après les données d'enquête.

6.5 Évaluation du rendement de l'étalonnage

Figure 7. Durée moyenne du congé de maternité, 1976 à 1996

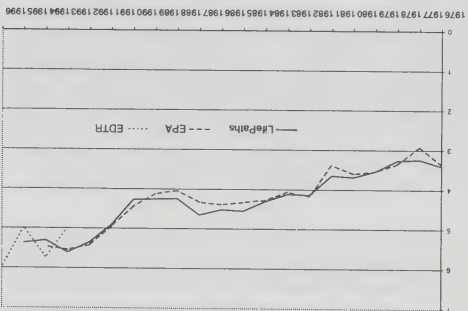
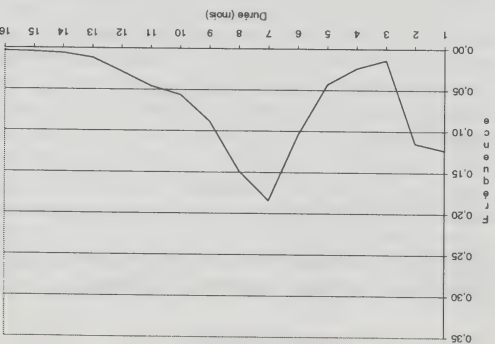


Figure 6. Données de l'EDTR : Distribution des durées des congés pour 1993 à 1996



La méthode d'étalonnage semble être très efficace dans le cas du modèle logit binaire. La tendance qui se dégage des données de l'EPA est bien reflétée par la simulation LifePaths. Dans le cas du modèle de survie, l'aspect clé des données de l'EPA est le déplacement abrupt du mode de la distribution des fréquences après 1990, déplacement qui semble être dû à l'introduction des prestations parentales. Les données simulées reflètent ce décalage. En outre, la durée moyenne du congé de maternité établie d'après la simulation concorde très étroitement avec les données de l'EPA.

Une divergence notable entre la simulation et les données de l'EPA est la hauteur du mode dans l'intervalle (3,4] mois dans la simulation LifePaths pour la période de 1982 à 1989. Cette divergence pourrait être due à l'effet de tendances caractérisant les valeurs des variables explicatives, variables que nous avons considérées comme étant stables. D'autres travaux seront nécessaires pour établir ce fait. Nous avons discuté d'une extension possible du modèle à la section 3.3.

6.4 Résultats de la simulation pour le modèle

étalonné de survie

Dans le cas du modèle semiparamétrique de survie, l'étalonnage consiste à rajuster tous les termes GAMMA, $i = 1, 2, \dots, 15$ de la section précédente conformément à (5.8) pour chacune des années de la période allant de 1975 à 1992. Nous intégrerons le modèle dans Lifepaths et exécutons la simulation.

Nous présentons la distribution des fréquences des durées simulées des congés de maternité et la comparons à la distribution correspondante des fréquences observées d'après les données de l'EPA. Aux fins de la présentation des résultats, nous avons calculé la moyenne des fréquences sur trois ans. Un aspect important de la courbe de distribution des fréquences est la variation abrupte apparemment liée à l'introduction des prestations parentales au moment de la mise en application du projet de loi C-21 à la fin de 1990. Comme les mères ayant présenté une demande de prestations pour maternité à l'époque de la mise en application de la loi avaient droit aux prestations parentales, les demandes enregistrées au début de 1990 représentent un mélange de régimes de prestations. Par conséquent, nous n'avons inclus l'année 1990 dans aucune des moyennes sur trois ans. Aux figures 2 et 3, nous utilisons des périodes de trois ans disjointes couvrant la période de 1976 à 1984. Pour équilibrer les périodes étudiées avant et après 1990 en nous servant des données disponibles, nous utilisons dans les figures 4 et 5 les périodes chevauchantes allant de 1985 à 1987, de 1987 à 1989, de 1991 à 1993 et de 1993 à 1995.

La distribution des durées des congés obtenue d'après les données de l'EDTR recueillies pour la période de 1993 à 1996 est présentée à la figure 6. Cette distribution peut être comparée aux données simulées présentées à la figure 4 pour la période allant de 1993 à 1995, puis jusqu'à aucun étalonnage n'a été appliqué après 1992.

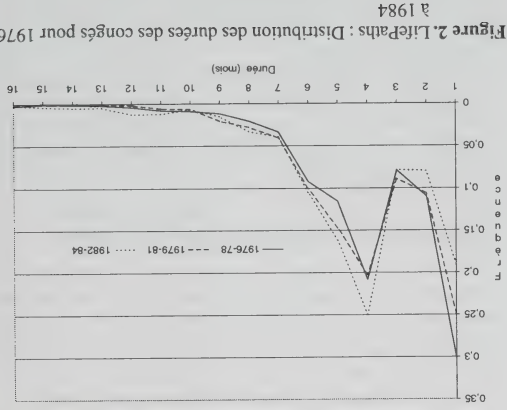


Figure 2. Lifepaths : Distribution des durées des congés pour 1976 à 1984

de niveau de scolarité n'ont aucun effet significatif sur le fait qu'une mère retourne au travail dans le mois. Toutefois, nous constatons que le fait de posséder un diplôme universitaire a un effet négatif significatif sur la fonction de hasard (effet positif sur la durée). La durée de l'emploi a un effet négatif significatif sur la fonction de hasard, ce qui reflète peut-être le lien entre cette variable et le droit à l'assurance-chômage ainsi que la sécurité de l'emploi.

Tableau 2

Estimations des paramètres du modèle de survie

Paramètre	Erreur-type	Valeur p
Durée de l'emploi (mois) /10	-0,030	0,0024
T-N.	0,195	0,6470
I.P.E.	0,307	0,5313
N-E.	0,173	0,253
N-B.	0,109	0,293
QUÉ	0,111	0,117
MAN	-0,402	0,253
SASK	-0,303	0,213
ALB	0,270	0,154
C.B.	-0,440	0,148
Travail autonome?	1,665	0,157
Âge	-0,253	0,041
Âge** 2 / 10	0,043	0,007
Premier enfant?	-0,301	0,090
< Diplôme d'études secondaires	0,508	0,206
Diplôme d'études secondaires	-0,124	0,125
Diplôme universitaire	-0,374	0,108
Conjoint employé?	0,109	0,151
Gammal	2,70	0,609
Gammal2	-1,136	0,816
Gammal3	-0,466	0,719
Gammal4	0,78	0,2232
Gammal5	6,215	0,0101
Log du rapport de variscblance = -1165,06		
Nombres d'observations = 3 411		
Les coefficients de pondération longitudinaux de l'EDTR pour l'année de la naissance, mis à l'échelle de sorte que leur somme soit égale à l'effectif de l'échantillon, sont appliqués aux observations.		

L'effet de la variable de premier enfant semble égale-

ment raisonnable. Pour une femme qui donne naissance à son premier enfant, la cote exprimant la possibilité de prendre un congé de maternité ne représente que 59 % de celle observée pour une femme qui a déjà des enfants, toutes les autres caractéristiques étant par ailleurs égales; autrement dit, les femmes qui deviennent mère pour la première fois sont plus susceptibles de cesser de travailler que celles qui ont déjà eu des enfants. Ces résultats pourraient tenir au fait que notre échantillon comprend les mères qui travaillent dans les quatre mois avant et après la naissance. Les femmes qui ont plus d'un enfant ont tendance à espacer les naissances d'au moins quelques années. Si elles sont employées juste avant une deuxième naissance ou avant les naissances subséquentes, elles auront déjà prouvé qu'elles sont retournées au travail après une absence dont la durée doit avoir été inférieure à l'intervalle entre les naissances. Cette observation permet d'exclure une tendance commune au retrait du marché du travail — par exemple, rester à la maison jusqu'à ce que les enfants aillent à l'école.

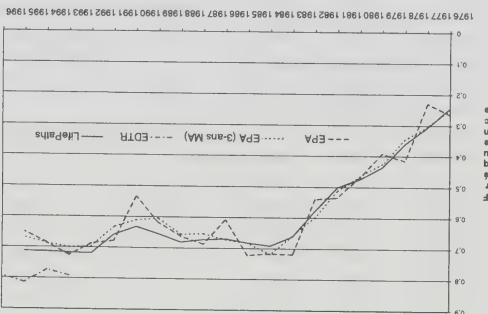
L'influence de l'âge est plus difficile à interpréter, puisqu'il s'agit de celui-ci sur le logarithme du rapport de cotes. Nous n'est pas linéaire. Si nous traçons le graphique du terme $-0,479 * \text{âge} - 0,0071 * \text{âge}^2$, nous constatons que, à mesure que l'âge augmente, le logarithme de la cote exprimant la chance de prendre un congé de maternité diminue jusqu'à un plateau où le maximum du logarithme commence par augmenter, mais que le taux de croissance diminue jusqu'à un plateau où le maximum du logarithme de la cote est atteint, à l'âge de 34 ans. Puisque le nombre de femmes qui donnent naissance à des enfants diminuent considérablement après cet âge, la diminution subséquente du logarithme pourrait ne pas être significative. Nous pourrions conjecturer que, chez les jeunes mères, le fait d'être, relativement parlant, plus âgée indique un attachement plus prononcé au marché du travail et, donc, une plus forte tendance à prendre un congé de maternité, alors que chez les mères plus âgées, qui ont passé l'étape de la première entrée sur le marché du travail, cet effet est moins prononcé. Cependant, les résultats ne sont sans doute pas suffisamment précis pour tirer des conclusions catégoriques à ce sujet.

6.2 Résultats des simulations pour le modèle logit binaire étalonné

L'exercice d'étalonnage consiste à ajuster le terme constant du modèle de la façon décrite par (5.2) pour chaque année, de 1975 à 1992. Nous n'ajustons pas le terme constant après 1992, en partie parce que les données de l'EPA ne révèlent aucune tendance marquée après 1992. Ensuite, nous intégrons le modèle dans LiFePaths et exécutons une simulation. Pour chaque année, de 1976 à 1995, la figure 1 montre la fréquence à laquelle est choisi le congé de maternité dans la simulation LiFePaths et la fréquence correspondante estimée d'après les données de l'EPA. Pour la période allant de 1993 à 1995, nous

6.3 Estimations pour le modèle de survie

Figure 1. Fréquence du choix du congé de maternité, 1976 à 1996



présentons aussi les estimations calculées d'après les données de l'EDTR. La simulation reflète la variation au fil du temps qui se dégage des données de l'EPA pour la période allant de 1976 à 1992. La simulation LiFePaths ne comporte aucune correction d'étalonnage après 1992, de sorte que les paramètres de base estimés d'après les données regroupées de l'EDTR pour 1993 à 1996 soient efficaces. La fréquence simulée est légèrement inférieure à la fréquence observée dans le cadre de l'EDTR durant cette période. Les deux sources possible d'erreurs sont le manque de souplesse de la spécification du modèle de choix binaire et les différences entre les estimations des variables explicatives basées sur les données de l'EDTR et celles fournies par le modèle LiFePaths.

Les estimations obtenues par application du modèle semiparamétrique de survie aux données de l'EDTR sont présentées au tableau 2. Comme dans le cas de l'estimation au moyen du modèle logit binaire, sont omises les variables indicatrices servant de catégorie de référence, c'est-à-dire « Ontario » pour la province de résidence et « Certaines études postsecondaires » pour le plus haut niveau de scolarité. Puisque la variable dépendante est la probabilité de retourner au travail, un coefficient de covariable positif indique une influence qui a tendance à raccourcir la durée du congé de maternité.

Les estimations des termes constants dans le prédicteur linéaire dépendant de la durée donné par (4.7) sont représentées dans le tableau 2 par GAMMA, $i = 1, 2, \dots, 15$. Ces estimations correspondent à l'effet de la fonction de hasard de base à laquelle est intégré l'effet de la durée. De nouveau, nous avons testé les variables de revenu personnel et de revenu familial et constaté que leur effet n'était pas significatif. Ces deux résultats, ainsi que l'importance de la variable de travail autonome en tant que prédicteur d'un retour précoce au travail concordent avec les résultats Marshall (1999). Selon ce dernier, les variables

Nous présentons les résultats de l'estimation pour la période de référence et les résultats de la simulation basée sur des estimations étalonnées des paramètres. Nous comparons les résultats de la simulation aux fréquences annuelles, calculées après les données d'enquête, du choix du congé parental dans les cas où le conjoint est distinctif des fréquentes années, calculées après les données d'enquête, du congé parental dans le cas où le conjoint est distinctif des fréquentes années, calculées après les données d'enquête, du congé parental dans le cas où le conjoint est distinctif des fréquentes années.

6. RÉSULTATS EMPIRIQUES

$$(5.7) \quad \cdot \left(\frac{[(i)_{o_1} x - 1] u_1}{[(i)_{o_1} x - 1] u_1} \right) u_1 + (i)_{o_1} x = (i)_{o_1} x$$

censuration à l'instant t (la censuration ne peut avoir lieu qu'à la fin des intervalles). Le nombre de mètres a été calculé d'après des dénombrements sur échantillon en appliquant le coefficient de pondération de l'EPA pour le mois durant lequel une nouvelle mère retourne au travail. La fonction de hasard empirique et l'estimateur correspondant de la fonction de survie découlant de la loi du produit des probabilités ont été étudiés par Kaplan et Meier (1958). L'utilisation de la fonction de hasard empirique dans l'équation (3.7) simultanément à l'équation (3.5) donne

6.1 Estimations dans le cas du modèle logit binaire

Les estimations obtenues par application du modèle logit aux données de l'ÉPTR sont présentées au tableau 1. Y sont omises les variables indicatrices formant les catégories de référence pour les variables utilisées dans le modèle, c'est-à-dire « Ontario » pour la province de résidence et « Certaines études postsecondaires » pour le niveau le plus élevé de scolarité. Nous avons testé les variables de revenu personnel et de revenu familial et constaté que leur effet n'était pas significatif, si bien que nous ne les avons pas incluses dans la régression. Les estimations peuvent être entachées d'un certain biais,

particulièrement celles de l'erreur-type, parce qu'on n'a tenu compte du plan d'échantillonnage complexe de l'EDTR qu'au moyen des coefficients de pondération appliqués au logarithme du rapport de vraisemblance.

L'effet positif significatif de la durée de l'emploi semble raisonnable pour plusieurs raisons. Une durée importante d'occupation d'un emploi pourrait indiquer que la femme commencée a accumulé un capital humain propre à l'entreprise et a atteint un certain niveau d'ancienneté. Elle

pourrait aussi acquiescer un attachement pressant du travail en général. Par ailleurs, le congé que l'entreprise accorderait à la femme avec la garantie qu'elle retrouvera son emploi par la suite sera d'autant plus long que l'employée aura de l'ancienneté. En outre, les garanties d'emploi offertes par les gouvernements provinciaux dépendent aussi de l'ancienneté professionnelle. Enfin une grande

ancienneté professionnelle signifie que la femme répondra aux exigences d'admissibilité aux prestations d'assurance-chômage (20 semaines d'emploi assuré). Nous avons testé dans le modèle une variable nominale indiquant que les critères d'admissibilité aux prestations d'assurance-chômage étaient satisfaisants et constaté que son effet était positif significatif au niveau de 5 %. Toutefois, nous n'avons pas inclus la variable dans le modèle, parce qu'à ce stade, nous ne sommes pas capables de modéliser les variations du programme d'assurance-chômage sous l'influence de covariables, à cause de l'incertitude de l'interprétation et de la forte corrélation avec la durée de l'emploi. Dans le cas de l'EPA, les travailleurs autonomes ne sont considérés comme ayant quitté le marché du travail que lorsqu'ils mettent fin à leur entreprises. Puisque prendre un congé signifie simplement ne pas mettre fin à l'entreprise, on devrait s'attendre à un effet positif significatif de la variable de travail autonome. Le fait d'avoir travaillé pour son propre compte avant la naissance augmente de 33,3 % la cote exprimant la chance de prendre un congé de maternité, ce qui représente l'effet le plus important observé pour une variable explicative.

Tableau 1

Estimation des paramètres du modèle logit binaire			
Paramètre	Estimation du coefficient	Contribution au rapport de cotes*	Erreur-type du coefficient
Valeur			p

Constante	T.-N.	I.-P.-E.	N.-B.	Qc	Man.	Sask.	Alb.	C.-B.	Durée de l'emploi (mois)/10	Travail autonome? 1,203	Âge (années) 0,479	< (Âge ²)/10 -0,071	> Diplôme d'études secondaires -0,702	Diplôme d'études secondaires -0,148	Diplôme secondaires -0,292	Premier enfant? universitaire
-6,432	-0,829	0,931	-0,207	-0,361	-0,490	-0,163	-0,200	-0,120	0,094	3,330	1,614	0,931	0,496	0,862	0,747	0,592
0,0318	0,2636	1,612	0,3992	0,7596	0,1437	0,3306	0,458	0,5379	0,300	0,418	0,199	0,033	0,357	0,276	0,229	0,192
0,0040	0,0040	0,0160	0,0296	0,0490	0,0003	0,0003	0,0003	0,0003	0,0003	0,0040	0,0160	0,0296	0,0490	0,5913	0,2027	0,0063

$$\text{Log du rapport de vraisemblance} = -381,553$$

Nombre d'observations = 835

Les coefficients de pondération

la naissance, mis à l'échelle de

de l'échantillon, sont appliqués

* Il s'agit de l'exponentielle du coefficient. Elle peut être interprétée comme étant la variation proportionnelle du rapport de cotes due à une variation unitaire de la variable indépendante correspondante.

4.3 Modèle semiparamétrique de survie : Avec décision concernant l'intervalle arrêt du travail-naissance

Dans notre application, la situation est compliquée du

travail et la naissance (intervalle arrêt de travail-naissance), ainsi que la fonction de retour au travail après un congé de maternité. Le modèle de l'intervalle arrêt du travail-naissance est estimé séparément, d'après des données de l'EDTR. L'examen de l'intervalle moyen pour chaque année de données de l'EPA indique que la durée de cet intervalle est restée assez stable au fil du temps, si bien que nous n'étalonnons pas le modèle. Néanmoins, il est nécessaire d'apporter une modification au modèle semi-paramétrique de survie afin d'intégrer le modèle distinct d'intervalle arrêt du travail-naissance. Nous pouvons pour cela supposer que la décision concernant l'intervalle arrêt de travail-naissance, qui tient éventuellement compte de l'état de santé, agit comme une contrainte sur la durée totale souhaitée du congé. Autrement dit, le modèle susmentionné s'appliquerait à la durée totale souhaitée, que l'on ne peut observer, et pourrait être appelée T^* .

Dans les cas où la durée souhaitée est plus courte que l'intervalle arrêt du travail-naissance, la mère pourrait retourner au travail aussitôt que possible après la naissance. Autrement dit, dans les cas où nous observons un intervalle arrêt de travail-naissance important (supérieur à un mois) et que la mère retourne au travail peu de temps après la naissance (dans le mois qui suit), tout ce que nous savons au sujet de la durée souhaitée est que

$$T^* \leq T$$

que l'observation ne commence. Pour ce genre d'observation, nous obtenons la contribution à la fonction de vraisemblance et à son logarithme par

$$(4.11) \quad L_i = 1 - \prod_{k_i} P(T^* \geq t | T^* \geq t - 1) = 1 - \prod_{k_i} \exp[-\exp(\eta_i(t))]$$

$$(4.12) \quad \ln(L_i) = \ln\{1 - \exp[-\sum_{k_i} \exp(\eta_i(t))]\}.$$

Malheureusement, l'expression du rapport de vraisemblance ne se simplifie pas comme l'expression correspondante pour les observations « censurées à droite ». Malgré cela, les expériences de Monte Carlo indiquent que l'estimation ne pose pas de difficulté même si les ensembles de données sont fortement censurés.

5. ETALONNAGE DES MODELES

Nous appliquons les coefficients de pondération longitudinaux de l'EDTR pour l'année de la naissance de l'enfant de la même façon que dans le cas de la forme élémentaire du modèle de survie.

Pour entamer la procédure d'étalonnage, nous devons inverser la fonction de distribution F donnée par l'équation (3.2) afin d'obtenir la fonction de lien g . Puis, nous appliquons l'équation (3.5) dans le cas du modèle logit et l'équation (3.7) dans le cas du modèle de survie.

5.1 Application du modèle logit binaire

Pour étalonner le modèle logit, nous devons d'abord inverser la fonction de distribution logistique dans l'équation (4.1) afin d'obtenir

$$(5.1) \quad \eta_i^* = g(\pi_i^*) = \ln \left(\frac{1 - \pi_i^*}{\pi_i^*} \right)$$

où g est la fonction logit bien connue. Nous pouvons alors appliquer les équations (3.5) et (5.1) pour obtenir

$$(5.2) \quad \eta_i^* = \eta_0^* + g(\pi_i^*) - g(\pi_0^*) = \eta_0^* + \ln \left(\frac{\pi_i^*/(1 - \pi_i^*)}{\pi_0^*/(1 - \pi_0^*)} \right)$$

où, pour $\tau < \tau_0$, chaque π_i^* représente la fréquence à laquelle est choisi le congé de maternité calculé d'après les données de l'EPA pour les congés de maternité commençant dans l'année τ , et π_0^* représente la fréquence calculée d'après les données de l'EDTR.

5.2 Extension au modèle de survie

Partant de l'équation (4.7) nous obtenons

$$(5.3) \quad \pi_i^*(t) = 1 - \exp[-\exp\{\eta_i^*(t)\}] = F\{\eta_i^*(t)\}$$

où

$$(5.4) \quad \eta_i^*(t) = \beta'x_i^*(t) + \gamma_i^*(t).$$

Ici, F est une distribution de valeurs extrêmes qu'il est facile d'inverser pour obtenir

$$(5.5) \quad \eta_i^*(t) = \ln[-1 - \ln[1 - \pi_i^*(t)]] = g(\pi_i^*(t)).$$

Pour l'étalonnage, nous pouvons utiliser l'équation (3.7) ainsi que les fréquences observées pour l'année τ représentée par le hasard empirique, ou ratio survenue/exposition, donné par

$$(5.6) \quad \pi_i^*(t) = d^*(t) / r^*(t)$$

où, pour les congés commençant à l'instant τ , $d^*(t)$ représente le nombre de mères dont le statut a changé dans l'intervalle $(t - 1, t]$ et $r^*(t)$ représente le nombre de mères observées pour la durée t , y compris celles éliminées par

Les équations pondérées de caractérisation (équations de score) sont

$$\frac{\partial \log L(\beta, \gamma)}{\partial \beta} = \sum_i w_i x_i y_i - \sum_i w_i x_i' F(\eta_i^*) = 0$$

$$\frac{\partial \log L(\beta, \gamma)}{\partial \gamma} = \sum_i w_i y_i - \sum_i w_i y_i' F(\eta_i^*) = 0. \quad (4.4)$$

Nous avons obtenu la solution, qui maximise le logarithme du rapport de vraisemblance, par itération de Newton-Raphson. Le modèle logit, qui est utilisé fréquemment par les statisticiens et les économétriciens, a fait l'objet de nombreux articles. Par exemple, consulter Chambless et Boyle (1985), Roberts, Rao et Kumar (1987) et Morel (1989).

4.2 Modèle semiparamétrique de survie : Forme élémentaire

Dans le cas où la mère choisit de prendre un congé de maternité, nous utilisons un modèle de survie pour décrire la durée du congé. La densité de probabilité de la distribution à une forme complexe, comme l'illustre les graphiques de la section 6.4. On observe un pic pour les durées inférieures à un mois et le mode semble correspondre au montant maximal des prestations spéciales d'assurance-chômage dont ont bénéficié les mères après 1990 (15 semaines de prestations de maternité, plus 10 semaines de prestations, plus une période d'attente de deux semaines). Pour commencer, nous avons estimé divers modèles entièrement paramétriques, y compris un modèle log-logistique de survie combiné à un modèle logit pour prédire les durées inférieures à un mois, mais nous n'avons pu obtenir un ajustement convenable. Pour résoudre ce problème, nous nous inspirons de Prentice et Gloeckler (1978), de Han et Hausman (1986) et de Meyer (1990) et nous estimons de façon non paramétrique les effets du temps sur la fonction de hasard de retour au travail. Cette dernière est exprimée sous forme d'une équation à hasards proportionnels :

$$\lambda_1^*(t) = \lambda_0^*(t) \exp \{ \beta' x_1^*(t) \} \quad (4.5)$$

où $\lambda_1^*(t)$ est la fonction de hasard de base, inconnue pour la durée t du congé et la période t , $x_1^*(t)$ est un vecteur de variables explicatives pour la mère i , et β est un vecteur de coefficients. Les données nous indiquent lequel des intervalles $[0, 1), [1, 2), [2, 3), \dots$ contient la durée du congé (dans notre cas l'unité est le mois) et le modèle peut être interprété comme un modèle de hasard en temps continu incomplètement observé sans restriction quant à la forme de la fonction de hasard de base. Si T_1^* est la durée du congé de la mère i à la période t , alors, pour $t = 1, 2, 3, \dots$, la probabilité que le congé dure jusqu'à la période t , étant donné qu'il a duré jusqu'à la période $t - 1$, peut être écrite sous la forme

$$P(T_1^* > t | T_1^* \geq t - 1) = \exp \left[- \int_{t-1}^t \lambda_1^*(u) du \right]$$

$$= \exp \left[- \beta' x_1^*(t) \int_{t-1}^t \lambda_0^*(u) du \right] \quad (4.6)$$

si nous supposons que $x_1^*(t)$ est constante sur l'intervalle entre $t - 1$ et t . Afin d'appliquer la théorie de la section 3, nous pouvons réécrire l'équation (4.6) sous la forme

$$1 - \pi_1^*(t) = P(T_1^* \geq t | T_1^* \geq t - 1)$$

$$= \exp \left[- \exp \{ \beta' x_1^*(t) + \gamma_1^*(t) \} \right] \quad (4.7)$$

$$= \exp \left[- \exp \{ \eta_1^*(t) \} \right]$$

où

$$\gamma_1^*(t) = \ln \left[\int_{t-1}^t \lambda_0^*(u) du \right]. \quad (4.8)$$

Nous pouvons censurer toute observation courante à partir d'une valeur élevée T de la durée. De nouveau, nous pouvons estimer les paramètres de base β et γ^0 au moyen du logarithme du rapport de vraisemblance $\ln L(\gamma^0, \beta)$. Comme nous référons systématiquement à des données recueillies pour la période de référence, dans la suite de la section 4, nous abandonnons les indices t_0 .

$$L(\gamma, \beta) = \prod_{N=1}^N [1 - \exp \{ - \exp(\eta_i(k_i^*)) \}]^{\delta_i}$$

$$\prod_{k_i=1}^{t_i} \exp \{ - \exp(\eta_i(t_i)) \} \quad (4.9)$$

où $\gamma = [\gamma(1), \gamma(2), \dots, \gamma(T)]'$, C_i est une période de censure, $\delta_i = 1$ si $T_i \leq C_i$ et 0 autrement, $k_i = \min(\text{int}(T_i), C_i)$. Par conséquent, le logarithme du rapport de vraisemblance est donné par

$$\ln L(\gamma, \beta) = \sum_{N=1}^N [\delta_i \ln [1 - \exp \{ - \exp(\eta_i(k_i^*)) \}]]$$

$$- \sum_{k_i=1}^{t_i} \exp(\eta_i(t_i)) \quad (4.10)$$

Les coefficients de pondération applicables aux mois durant lesquels un enfant est observé pour la première fois sont mis à l'échelle de sorte que leur somme corresponde à l'effectif de l'échantillon, puis utilisés pour pondérer les termes de la fonction de vraisemblance et ses dérivées. La fonction de pondérée de Newton proposé par Bryden, Fletcher, Goldfarb et Shanno (BFGS), selon une méthode d'application fondée sur Dennis et Schnabel (1983).

où $\hat{\pi}^t$ est une estimation de $E(X^t)$. Si nous utilisons les données de l'événement à la période 1 (en prenant chaque coefficient de pondération du mois pour lequel on observe un enfant pour la première fois). Pour justifier cette méthode, nous utilisons l'équation (3.4) et supposons une approximation

$$E\{g(\pi^t(X^t))\} - E\{g(\pi^0(X^0))\} \approx g(E\{\pi^t(X^t)\})$$

$$-g(E\{\pi^0(X^0)\}) \quad (3.6)$$

En raison de l'inégalité de Jensen, les résultats seront inexacts dans les régions où g est convexe ou concave. Néanmoins, si, localement, g s'approche d'une fonction linéaire dans les régions où $\pi^0(X^0)$ et $\pi^0(X^0)$ sont concentrées, alors (3.6) pourrait être relativement exacte. Le fait que g possède un point d'inflexion à 0,5 pourrait faciliter l'approximation lorsque les probabilités sont dispersées autour de cette valeur.

Heureusement, nous pouvons tester l'adéquation de l'estimateur par simulation du modèle estimé dans LifePaths et comparaison des fréquences prévues de l'événement aux fréquences pondérées correspondantes calculées d'après les données. Les résultats indiquent que le modèle est suffisamment adéquat pour notre application.

3.2 Application à l'analyse de survie

Nous montrerons à la section 5.2 que la méthode décrite plus haut peut être étendue au modèle semiparamétrique de survie grâce à l'ajout d'un indice t représentant la durée dans l'état courant, si bien que (3.5), devient

$$\eta^t(t) = \eta^0(t) + g(\pi^t(t)) - g(\pi^0(t)) \quad (3.7)$$

où $\pi^t(t)$ représente la fonction de hasard empirique.

3.3 Tendances qui caractérisent les variables indépendantes

Nous pouvons améliorer la méthode d'échantillonnage en tenant compte de l'évolution des caractéristiques observées. Comme nous l'avons mentionné à la section 2.3, cette correction serait envisagée si d'autres éléments du modèle LifePaths étaient à un stade de développement plus poussé. Pour cela, nous relâchons l'hypothèse selon laquelle les vecteurs aléatoires X^t sont identiquement distribués. L'équation (3.4) devient alors

$$E\{\eta^t(X^t) - \eta^0(X^0)\} = \eta^t - \eta^0 + \beta' \{E(X^t) - E(X^0)\}$$

$$= E\{g(\pi^t(X^t))\}$$

$$- E\{g(\pi^0(X^0))\} \quad (3.8)$$

D'après cette équation, nous pourrions estimer η^t par

$$\hat{\eta}^t = \hat{\eta}^0 + g(\hat{\pi}^t) - \beta'(\hat{\pi}^t - \hat{\pi}^0) \quad (3.9)$$

où $\hat{\pi}^t$ est le vecteur des valeurs moyennes des caractéristiques à la période t . Naturellement, il pourrait être impossible d'obtenir toutes les valeurs moyennes à partir de la

même source de données. La méthode s'étendrait au modèle de survie de la même façon que (3.7) pour donner

$$\eta^t(t) = \eta^0(t) + g(\pi^t(t)) - g(\pi^0(t))$$

$$- \beta'(\pi^t(t) - \pi^0(t)) \quad (3.10)$$

4. MODÈLES ET ESTIMATION DES PARAMÈTRES DE BASE

Comme nous l'expliquons à la section 3.1, les paramètres de base β et η^0 sont estimés par la méthode du maximum de vraisemblance au moyen de données recueillies pour la période t_0 . Nous utilisons les données de l'EDTR sur les congés de maternité à compter de la période allant de 1993 à 1996 (notre période de référence t_0). Nous n'essayons pas d'estimer la variation annuelle du terme constant γ durant cette période.

4.1 Modèle logit binaire

Nous adoptons le modèle logit pour représenter le choix fait par une mère entre le congé de maternité et le retrait du marché du travail. À partir d'ici, nous adoptons une notation économique plus conventionnelle et utilisons l'indice t pour représenter une variable aléatoire ou un résultat associé à un individu i . Nous supposons que la variable aléatoire X_i^t prend la valeur 0 ou 1, $X_i^t = 1$ indiquant que la nouvelle mère i à laquelle est associé le vecteur de caractéristiques x_i^t à la période t choisit de prendre un congé de maternité, à condition qu'elle ait été employée, et que

$$\pi_i^t = P(Y_i^t = 1) = F(\eta_i^t) = \frac{1 + \exp(\eta_i^t)}{\exp(\eta_i^t)} \quad (4.1)$$

où $\eta_i^t = \beta'x_i^t + \gamma^t$ est le prédicteur linéaire de l'équation (3.1) et F est la fonction de distribution logistique. Nous estimons les paramètres de base β et η^0 au moyen de N observations provenant de l'EDTR en maximisant le logarithme du rapport de vraisemblance $\ln L(\beta, \gamma^0)$ où

$$L(\beta, \gamma^t) = P(X^1 = \gamma^1, X^2 = \gamma^2, \dots, X^N = \gamma^N) \\ = \prod_{\gamma^t=0}^1 [1 - F(\eta_i^t)] \prod_{\gamma^t=1}^1 F(\eta_i^t)$$

$$= \prod_{i=1}^I [F(\eta_i^1)]^{\gamma_i^1} [1 - F(\eta_i^1)]^{1-\gamma_i^1} \quad (4.2)$$

et

$$\ln L(\beta, \gamma^t) = \sum_{i=1}^I \{ \gamma_i^t \ln F(\eta_i^t) + (1 - \gamma_i^t) \ln [1 - F(\eta_i^t)] \} \quad (4.3)$$

Les coefficients de pondération longitudinaux de l'EDTR calculés pour l'année de la naissance de l'enfant sont mis à l'échelle de sorte que leur somme soit égale à l'effectif de l'échantillon, puis utilisés pour pondérer les termes du logarithme du rapport de vraisemblance et de ses dérivées.

utilisons des méthodes assez simples dont nous discutons aux sections 4 et 5.

2.3 Le problème de l'étalonnage

Notre problème d'étalonnage a pour contexte un modèle représentant les femmes qui font le choix entre le retrait du

marché du travail ou le départ en congé de maternité et qui, si elles choisissent ce dernier, décident de la durée de ce

congé. La première décision est représentée par un modèle

logit binaire et la deuxième, par un modèle semiparamé-

trique de survie. Ces modèles incluent l'un et l'autre un

vecteur de variables explicatives et de paramètres connexes.

Dans le modèle de simulation LifeFahs, les décisions sont

prises dans le cadre de l'événement représentant le choix

ayant trait au congé de maternité, choix qui a toujours lieu

au milieu de la grossesse. Les données de l'EDTR con-

viennent assez bien à l'estimation des paramètres de base de

ces deux modèles. Cependant, puisqu'un des objectifs

d'évolution chronologique des processus socioécono-

miques, il a fallu étalonner les estimations des paramètres

fondées sur les données de l'EDTR en fonction des esti-

mations annuelles de la moyenne des variables dépendantes

calculées d'après les données de l'EPA.

Pour traiter ce problème, nous supposons, pour deux

raisons, que les caractéristiques observées de la population

sont stables. En premier lieu LifeFahs est en cours d'éta-

lisation et l'exercice d'étalonnage dont nous présentons les

résultats a été réalisé à un stade où d'autres éléments du

modèle destinés à prédire ces caractéristiques faisaient

l'objet d'une révision approfondie. Dans la section 3.3,

nous effleurons la question des conséquences de l'évolution

des caractéristiques de la population. En deuxième lieu,

nous supposons que l'évolution systématique des résultats

observés entre les périodes de référence est due principale-

3. MÉTHODE D'ÉTALONNAGE

ont épuisé les prestations d'assurance-chômage.

À la présente section, nous présentons la méthode sous

forme abstraite, afin de préciser les hypothèses, d'élaborer

la notation et de montrer la similitude entre l'application au

3.1 Application au choix binaire

choix binaire et à l'analyse de survie.

Le modèle de base de la méthode d'étalonnage a trait au

choix binaire. Puisque notre objectif principal n'est pas

d'observer les changements dans la population, nous

simplifions l'analyse en supposant que les variables expli-

catives ou les caractéristiques individuelles durant la

période τ sont représentées par une série de vecteurs

aléatoires indépendants et indépendamment distribués X^t .

Nous recommandons qu'il s'agit d'une hypothèse assez

forte. Néanmoins, pour des raisons dont nous discutons à la

section 2.3, nous l'utilisons pour nos travaux empiriques.

Nous montrons à la section 3.3 qu'il est assez simple

d'étendre la théorie afin d'intégrer les tendances observées

pour les variables dépendantes.

Considérons un prédicteur linéaire donné par

(3.1) $\eta^t(x) = \beta'x + \gamma^t$

où β est un vecteur de coefficients constant au cours du

temps, x est un résultat possible de X^t et γ^t représente un

paramètre particulier à la période τ . À noter que x ne

contient aucun « terme constant ». Posons que X^t est une

variable aléatoire, à distribution conjointe avec X^t , qui

prend la valeur 1 si un événement a lieu et 0 s'il n'a pas

lieu. Supposons que la probabilité de l'événement,

conditionnelle aux caractéristiques x , est donnée par

(3.2) $E(X^t | X^t = x) = \pi^t(x) = F(\eta^t(x))$

où nous posons que F est une fonction de distribution

continue. Cette fonction sera alors bornée par les

valeurs zéro et un, et possèdera une fonction inverse g , telle

que

(3.3) $\eta^t(x) = g(\pi^t(x))$.

Dans le contexte des modèles linéaires généralisés, g est

appelée fonction de lien. Nous commençons par calculer les

estimations du maximum de vraisemblance des paramètres

de base β et γ^0 à l'aide des données recueillies pour la

période de référence τ_0 (dans notre cas, il s'agit de la

période pour laquelle les données de l'EDTR sont

disponibles). Naturellement, ces données doivent inclure les

variables qui correspondent aux résultats de X^t ainsi que

des données provenant de l'EPA pour chaque période, nous

Puisque nous ne disposons, sur les résultats de X^t , que des

estimons les termes γ^t comme suit

(3.5)

$\gamma^t = \gamma^0 + g(\eta^t) - g(\eta^0)$

(3.4) $-E\{g(\pi^0(X^0))\}.$

$E\{\eta^t(X^t) - \eta^0(X^0)\} = \gamma^t - \gamma^0 = E\{g(\pi^t(X^t)) -$

2.2 Sources des données

Les paramètres de base du modèle du congé de maternité ont été calculés d'après les données de l'EDTR couvrant les congés de maternité à compter de la période allant de 1993 à 1996. L'utilisation de données recueillies pour 1997 nous a permis de suivre la plupart des congés de maternité jusqu'à leur achèvement plutôt que d'utiliser des données fortement censurées. L'EDTR est une enquête-ménage conçue pour permettre l'analyse longitudinale et transversale des situations financière et professionnelle des individus. L'enquête, qui a été lancée en 1993, suit les mêmes personnes pendant six ans, un nouveau groupe de rotation étant créé tous les trois ans. Chaque groupe de rotation compte 15 000 ménages regroupant 30 000 adultes. L'enquête nous fournit les données sur le mois de naissance de l'enfant, nous donne des mensuelles sur la situation d'activité et sur un riche ensemble de variables explicatives, y compris la durée de l'emploi, un indicateur du travail autonome, le rang de naissance de l'enfant, l'existence d'un conjoint employé, la province de résidence, le niveau de scolarité et l'âge. Nous pouvons aussi déterminer si une mère qui a quitté un emploi dans les quatre mois avant un accouchement a repris le même emploi dans les 16 mois qui suivent. Ces spécifications, que nous utilisons comme définition pratique du congé de maternité, deviennent notre unité d'analyse avec une légère extension afin d'inclure un pourcent de cas où une mère a repris un emploi différent après un état d'activité correspondant à une absence au cours du mois précédent. En utilisant cette unité d'analyse, nous obtenons un échantillon de 835 naissances. Comme nous le montrons à la section 6, cet échantillon est de taille suffisante pour dégager certains facteurs explicatifs importants. Plus précisément, nous constatons que l'effet de plusieurs facteurs est significatif au niveau de confiance de 95 %. Cet échantillon contient environ 730 mères distinctes, qui représentent plus de 87 % de l'échantillon de naissances. Autrement dit, il existera une certaine corrélation entre les observations, parce que certaines mères ont pris deux congés de maternité ou plus durant la période de référence, mais nous estimons que cette corrélation n'est pas suffisamment forte pour justifier l'utilisation d'outils statistiques spéciaux.

L'EPA est une enquête-ménage mensuelle qui se concentre sur la situation d'activité, mais qui fournit aussi des données sur plusieurs caractéristiques démographiques. Normalement, les données de l'enquête servent exclusivement à l'analyse transversale. Cependant, pour le projet LiFePath, nous avons créé un fichier couvrant la période de 1976 à 1995 afin de suivre les individus à mesure qu'ils passent par les six groupes de rotation mensuels de l'enquête, et d'obtenir ainsi une fenêtre sur six mois de l'activité de chaque individu sur le marché du travail. Comme le nombre d'enfants et leur âge est enregistré chaque mois, il est possible d'observer l'arrivée d'un nouvel enfant. Puisque nous utilisons toutes les enquêtes

Dans la fenêtre de six mois de l'EPA, nous notons la situation d'activité d'une nouvelle mère au moment où la présence de l'enfant est déclarée pour la première fois. Cette variable est la clé de l'estimation de la probabilité que la mère choisisse un congé de maternité plutôt qu'un retrait du marché du travail. Nous commençons par considérer $P(E)$, c'est-à-dire la proportion de mères de cette catégorie qui sont employées. Si la mère est « employée, au travail », nous supposons qu'elle s'est absente brièvement de son travail, c'est-à-dire moins d'un mois. Si elle est « employée, absente du travail », il se peut qu'elle ait choisi de prendre un congé de maternité, puis de reprendre son emploi après ce congé. Cependant, il n'en n'est pas toujours ainsi. Une nouvelle mère que nous considérons comme employée et absente (EA) peut plus tard quitter son emploi et passer à l'état de non-employée (NE). Pour tenir compte de cette situation, considérons les mères ayant un enfant de moins d'un an observées dans une fenêtre. Nous calculons la proportion $P(EA - NE)$ de transitions de l'état d'« employée, absente du travail » à l'état de « non-employée ». Nous estimons aussi la proportion $P(NE - AE)$ de mères qui retournent à leur ancien emploi (AE) après avoir quitté le marché du travail. Nous calculons cette estimation d'après les données sur les mères ayant un jeune enfant qui font la transition d'un état de non-employée à un emploi pour lequel la date de début est antérieure au mois précédent. Notre estimation de la probabilité de choisir le congé de maternité devient maintenant $P(E) - (P(EA - NE) + P(NE - AE))$. Il est également possible d'observer des mères ayant un enfant de moins d'un an qui passent de l'état d'« employée, absente du travail » pour des raisons personnelles ou familiales à l'état d'« employée, au travail ». Nous utilisons cette transition comme approximation du retour au travail après un congé de maternité. Puisque la durée de l'absence est déclarée le mois précédent, cet élément de donnée est essentiel à l'étalonnage du modèle de survie.

La discussion qui précède illustre la faiblesse des données de l'EPA comparativement à celles de l'EDTR dans le cadre d'une étude du congé de maternité. L'EPA fournit non seulement un moins grand nombre de variables explicatives que l'EDTR, mais elle nous oblige aussi à accepter des approximations pour les variables dépendantes. Néanmoins, nous avons besoin de la profondeur chronologique des données de l'EPA. Cette relation entre les ensembles de données constitue le contexte du problème

l'EPA et l'EPA ont toutes deux un plan d'échantillonnage complexe comprenant une stratification détaillée et des méthodes complexes de calcul des coefficients de pondération des observations. Nous nous servons systématiquement des coefficients de pondération aussi bien pour le calcul des estimations que pour celui des fréquences. Nous

2. CONTEXTE DU PROBLÈME

Pour placer le problème dans son contexte, nous commençons par donner un aperçu de la structure du modèle LifePaths, une brève description des sources de données utilisées et un exposé de la façon dont s'est posé le problème de l'étalonnage.

2.1 Structure du modèle LifePaths

Le modèle LifePaths permet de simuler le cycle de vie d'individus sous forme d'une série d'événements qui modifient l'ensemble de « variables d'état » décrivant les caractéristiques démographiques, sociales et économiques individuelles. Pour tout individu, des périodes d'attente sont associées à chaque événement possible, avec la possibilité d'être infinies. Les périodes d'attente peuvent aussi dépendre conditionnellement des valeurs des variables d'état. Le type d'événement pour lequel la période d'attente est la plus courte est celui qui se produit (c'est-à-dire que le modèle active les fonctions associées à cet événement). La modification de toute variable d'état au moment où a lieu un événement peut mener à la production de nouvelles périodes d'attente pour d'autres événements.

Pour initialiser un cas, le modèle LifePaths attribue de façon aléatoire le sexe, la province de résidence, l'âge au moment de l'immigration et l'année de naissance de l'individu « dominant » (L'année de naissance peut varier de 1892 à 2051. Les hypothèses concernant la mortalité et l'immigration sont conçues de façon à reproduire les structures provinciales âge-sexe. Lorsqu'un individu dominant se marie, établit une union de fait ou a un enfant, le modèle crée un individu non dominant ayant les caractéristiques appropriées qui est relié à l'individu dominant et fait alors partie du cas. Une fois créés, les individus non dominants subissent les mêmes événements éventuels que les individus dominants. Cependant, comme leur raison d'être est de compléter le profil de l'acteur dominant, les données qui les concernent sont généralement filtrées de tous les rapports tabulaires.

À l'heure actuelle, LifePaths comprend des modèles de fécondité, de mortalité, de nuptialité (y compris les unions de fait), de cheminement dans la formation, de carrière au sein de la population active, de congé de maternité, de nombre d'heures de travail, de revenu, d'impôt et de transferts. Le modèle des carrières au sein de la population active décrit les transitions entre les états d'« employé », de « travailleur autonome » et de « non-emploi ». Il comprend aussi un modèle du départ à la retraite et du travail d'étudiant. Le modèle de cheminement dans la formation au secondaire et postsecondaire au niveau provincial est bien au point et fortement développé.

d'une gamme variée d'analyses de politiques et de travaux de recherche stratégique. À titre d'exemple, mentionnons l'analyse de la politique canadienne de prêts aux étudiants (étude réalisée pour le compte de Développement des ressources humaines Canada et du gouvernement de l'Ontario), l'étude du retour aux études (Appleby, Boothby, Rouleau et Rowe, 1999), l'étude de l'utilisation du temps (Wolfson et Rowe, 1996; Wolfson 1997; Wolfson et Rowe, 1998), l'étude du régime de transfert d'impôt et des pensions (Wolfson, Rowe, Gribble et Lin 1998; Wolfson et Rowe 1998b) et l'étude des carrières au sein de la population active (Rowe et Lin 1999). En outre, la nécessité de produire des données pour LifePaths a donné lieu à de nouveaux travaux de recherche portant, par exemple, sur les carrières dans l'enseignement (Chen et Oderkirk 1998; Rowe et Chen 1997; Plager et Chen 1999) et sur la correction des gains (Chen et Rowe 1999).

Le modèle LifePaths est conçu pour intégrer les renseignements socioéconomiques provenant de toutes les sources pertinentes disponibles à Statistique Canada. Aussi sa construction a-t-elle motivé l'étude de méthodes permettant d'exploiter des sources multiples de données. L'intégration d'un modèle estime dans LifePaths est un outil puissant si l'on veut faire, d'après le modèle, des inférences que l'on peut comparer à des renseignements provenant d'autres sources. Par exemple, Rowe et Lin (1999) ont obtenu les données sur la durée d'occupation (en emploi par simulation au moyen d'un modèle estime d'après des données longitudinales couvrant une courte période, puis par comparaison des résultats aux données d'une enquête transversale. Nous décrivons ici l'un des volets des efforts déployés en vue de créer un outil qui fournira le maximum d'information pouvant être extraite des sources de données de Statistique Canada.

La présentation de l'article vise à illustrer la façon dont les problèmes techniques sont souvent identifiés durant l'élaboration du modèle LifePaths et comment leurs solutions sont intégrées dans le processus de développement du modèle. Par conséquent, nous fournissons une quantité assez importante de renseignements généraux sur certaines questions connexes. À la section 2, nous décrivons dans les grandes lignes le contexte du problème d'étalonnage et à la section 3, nous présentons la théorie qui sous-tend notre solution, en mentionnant certaines extensions possibles qui pourraient faire l'objet de futurs travaux. À la section 4, nous décrivons les modèles auxquels elle sera appliquée, y compris certains détails concernant l'application de leur paramètres pour la période de référence. Puis, à la section 5, nous décrivons l'application de la méthode d'étalonnage à ces modèles. À la section 6, nous présentons nos résultats empiriques et nous en discutons, puis, à la section 7, nous présentons certaines conclusions générales.

Étalonnage des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie

IAN CAHILL & EDWARD J. CHEN

RÉSUMÉ

Nous élaborons une méthode d'exploitation des données provenant de plusieurs enquêtes et périodes de référence grâce à l'établissement de paramètres estimés des modèles logit de choix binaire et des modèles socioéconomiques que constituent l'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada et l'horizon temporel de l'Enquête sur la population active (EPA) conjugué au suivi des individus lors de chaque interview durant les six mois où ils font partie de l'échantillon de l'enquête. Nous démontrons comment la méthode peut être appliquée à l'aide du module sur le congé de maternité du projet de microsimulation dynamique LifeAbas de Statistique Canada. Le choix consistant à prendre un congé de maternité plutôt que d'arrêter de travailler est spécifique à nos forme de modèle logit binaire, tandis que la durée du congé est spécifique à nous forme de modèle semiparamétrique de survie à risques proportionnels comprenant des covariables ainsi qu'une fonction de hasard de base qui peut varier chaque mois. Nous estimons d'abord les deux modèles par la méthode du maximum de vraisemblance au moyen des données regroupées de l'EDTR sur les congés de maternité à compter de la période de 1993 à 1996, puis nous les réajustons en fonction des estimations annuelles calculées d'après les données de l'EPA recueillies pour la période de 1976 à 1992. Dans le cas du modèle logit, nous ajustons le prédicteur linéaire d'après une estimation du logarithme de la cote exprimant la chance (log-odds) basée sur les données de EPA. Pour le modèle de survie, nous utilisons un estimateur de Kaplan-Meier de la fonction de hasard calculé d'après les données de l'EPA pour rajuster la valeur prévisible de la fonction de hasard dans le modèle semiparamétrique.

MOTS CLÉS : Microsimulation; étalonnage; modèles semiparamétriques de survie; logit binaire.

1. INTRODUCTION

Les chercheurs fondent souvent leurs modèles économétriques sur les données d'une enquête réalisée sur une courte période. Le cas échéant, il peut être souhaitable d'intégrer des renseignements provenant d'une source complémentaire couvrant une période plus longue, même si des mesures ne sont disponibles que pour la variable dépendante. Pour une classe générale de modèles non linéaires, nous mettons au point une méthode simple d'estimation des paramètres calculées d'après les données d'une enquête riche en variables explicatives en fonction des renseignements fournis par une enquête dont l'horizon temporel est considérable. L'un des objectifs principaux est de faire concorder les prévisions du modèle avec l'information provenant de la source secondaire de données. Nous décrivons d'abord l'application de la méthode à un modèle logit simple de choix binaire, puis à un modèle semiparamétrique de survie. Puisqu'il peut être interprété comme une série de choix binaires, tout en étant interprété comme un modèle temporel continu incomplètement observé, le modèle de survie représente une généralisation naturelle de la première application.

Nous procédons, pour illustrer la méthode, à une étude du congé de maternité. L'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada fournit

des données sur la fréquence à laquelle le congé de maternité est choisi de préférence au retrait du marché du travail et sur la durée du congé de maternité, ainsi que sur un riche ensemble de variables explicatives. Par conséquent, nous utilisons les données de cette enquête pour estimer les effets des variables explicatives sur la fréquence (modèle logit) et sur la probabilité de retourner au travail (modèle de survie). L'enquête sur la population active (EPA) réalisée par Statistique Canada fournit des approximations raisonnables de la fréquence et de la durée du congé de maternité remontant jusqu'à 1976. Par conséquent, les paramètres estimés d'après les données de l'EDTR sont étalonnées en fonction des estimations calculées d'après les données de l'EPA de la fréquence et de la probabilité de retour au travail pour la période allant de 1976 à 1992, période d'antériorité à celle pour laquelle existent des données de l'EDTR.

BIBLIOGRAPHIE

- processus d'imputation asynchrone (celui du recensement renové) par celle du processus d'imputation synchrone (proche du modèle de Särndal).
- Cette approche a été testée sur les petites et moyennes communes de Rhône-Alpes, pour lesquelles les groupes de rotation, la Taxe d'Habitation (TH90) et le Recensement général de la population de 1990 (RG90) sont disponibles (Kaufmann 2000). La méthode donne de bons résultats pour les variables bien corrélées avec la TH; les résultats indiquent aussi qu'une source de données administratives proches des variables décrivant les personnes sera nécessaire afin de maintenir les erreurs du modèle à un niveau tolérable.
- #### 4. TRAVAUX EN COURS
- Les travaux de méthodologie autour de la rénovation du recensement sont loin d'être complétés. Au nombre des chantiers ouverts, notons
- l'établissement des règles de passage de seuil, les problèmes d'oscillation autour du seuil des 10 000 habitants et le calcul des populations légales;
 - la sensibilité des bornes de strate en grande commune et leur robustesse dans le temps;
 - la mise à jour et la maintenance des bases de sondage et des échantillons, en particulier, les ajustements éventuels suite aux passages de seuil et l'incorporation de nouveaux objets dans les groupes de rotation;
 - l'imputation massive et la synthèse, tant les modèles que leur précision;
 - l'estimation de la précision des estimateurs; et la collecte auprès des populations mobiles.
- SARNDAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. *Recueil : Symposium 90. Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.
- BORCHSENIUS, L. (2000). From a Conventional to a Register-based Census of Population. Les Recensements après 2001. Séminaire Eurostat-INSEE, Paris.
- DEVILLE, J.C., et JACOD, M. (1996). Replacing the Traditional French Census by a Large Scale Continuous Population Survey. *Annual Research Conference Proceedings*, USBC, Washington.
- DEVILLE, J.C., et TILLÉ, Y. (1999). *Balanced Sampling by Means of the Cube Method*. CREST-ENSAI, document interne, soumis pour publication.
- DEVILLE, J.C., et TILLÉ, Y. (2000). Échantillonnage équilibré par la méthode du cube et estimation de variance. *Journées de Méthodologie*, décembre 2000, INSEE, Paris.
- HORVITZ, D.G. (1986). Statement to the Subcommittee on Census and Population. Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.
- KAUFFMANN, B. (2000). *Estimations annuelles dans la rénovation du recensement de la population*. Note de travail interne, Département de la démographie, INSEE.
- KISH, L. (1981). Population Counts from Cumulated Samples. Congressional Research Service. *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.
- KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 1, 63-71, Statistique Canada, Ottawa.
- LAIHONEN, A. (2000). 2001 Round Population Censuses in Europe. *Les Recensements après 2001*, Séminaire Eurostat-INSEE, Paris.
- NATHAN, G. (2001). Models for combining longitudinal data from administrative sources and panel surveys. Présentation invitée, ISI, Séoul, août 2001.
- SÄRNDAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. *Recueil : Symposium 90. Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.

3.5 Imprecision due à la synthèse

On a montré, à la section 3.2, comment la production des estimations par synthèse utilise l'information amassée : d'abord une extrapolation pour un « vieux » recensement, l'utilisation directe des résultats du recensement pour un troisième groupe de rotation (disons III); enfin, la combinaison des extrapolations et rétroprojections pour caler les deux derniers groupes (disons IV et V).

Cette synthèse peut être formalisée sous l'angle d'un modèle de non-réponse (Sæmål 1990) : la campagne annuelle s'apparente alors à un sondage à 100 % qui subirait 80 % de non-réponse, laquelle est palliée par le recours à l'imputation par le ratio. Si l'échantillon complet est noté s , les répondants sont notés r et les non-répondants sont notés $s-r$, on peut écrire

$$y_k = \begin{cases} x_k & \text{si } k \in r \\ \beta x_k & \text{si } k \in s-r \end{cases} \text{ avec } \beta = \frac{\bar{y}_r}{\bar{x}_r}$$

C'est-à-dire que le modèle d'imputation est

$$\begin{cases} y_k = \beta x_k + \varepsilon_k \\ E(\varepsilon_k) = 0 \\ V(\varepsilon_k) = \sigma^2 x_k \end{cases}$$

où les erreurs ε_k sont non corrélées. Avec un tel modèle d'imputation, sous sondage aléatoire simple,

$$y_k = \frac{n}{N} \sum y_k = \frac{n}{N} \left\{ \sum_r y_k + \sum_{s-r} \beta x_k \right\} = \dots = N \frac{\bar{x}_r}{\bar{y}_r} \bar{x}_s$$

L'incertitude autour de l'estimation avec imputation dépend des aléas de sondage et de la qualité du modèle d'imputation ξ :

$$(y - Y) = (Y - Y) + (Y - Y)$$
$$\text{incertitude} = \text{incertitude du sondage} + \text{incertitude totale}$$

Si on suppose que l'imputation se fait sans biais :

$$E \xi_s E_s E_r (Y - Y) = 0$$

on a,

$$V^{\text{totale}} = E \xi_s E_s E_r (Y - Y)^2 + E \xi_s E_s E_r (Y - Y)^2$$
$$= E \xi_s V_s + E_s E_r V_\xi$$
$$V^{\text{totale}} = V^{\text{échantillon}} + V^{\text{imputation}}$$

en supposant que le sondage et la réponse sont indépendants du mécanisme d'imputation.

L'utilisation de données imputées comme si elles avaient été observées dans le calcul de l'estimation de V_s même à une sous-estimation de $V^{\text{échantillon}}$. En espérance,

$$E_\xi (V_s - V_s) = V^{\text{diff}}$$

Pour les estimateurs de ces variances, Sæmål montre que l'on obtient

$$V^{\text{sondage}} = N^2 \left(\frac{1}{1} \sum e_k^2 + C_0 \sigma^2 \right)$$
$$V^{\text{imputation}} = N^2 \left(\frac{1}{1} \sum e_k^2 + C_0 \sigma^2 \right)$$

avec $A = \bar{x}_{s-r} / \bar{x}_r$, qu'on peut comprendre comme un effet de sélection des répondants. On remarque que si $x_k = 1$, alors, on obtient un sondage à deux phases de taille m parmi n et n parmi N . De plus, si $s = r$, $V^{\text{totale}} = V^{\text{sondage}}$.

Dans le modèle de Sæmål, les x (données administratives) et y (données censitaires) sont contemporains; à tout le moins, on aura observé certains des y . En reprenant la structure développée à la section précédente, on aurait :

Année A-2	
y_k	x_k
y_k	x_k

m répondants (Groupe III)

n-m imputations (autres groupes)

Dans l'application du RRP, tout n 'est pas synchrone :

... A-4	x_{A-4}^I	x_{A-4}^{II}	x_{A-4}^{III}	x_{A-4}^{IV}	x_{A-4}^V
A-3	x_{A-3}^I	x_{A-3}^{II}	x_{A-3}^{III}	x_{A-3}^{IV}	x_{A-3}^V
A-2	x_{A-2}^I	x_{A-2}^{II}	x_{A-2}^{III}	x_{A-2}^{IV}	x_{A-2}^V
A-1					
A					

De plus, pour les groupes IV et V, l'estimation synthétique pour A-2 pourrait profiter des informations recueillies durant la campagne de l'année A-1 (respectivement A); en effet, il serait possible de calculer des facteurs d'ajustement par rapport au plus récent recensement et rétropolier sur la période intercensitaire. Par exemple, pour une commune D du groupe IV, on peut faire :

$$\Theta_1 = R_{A-6}^{A-2} \times \frac{\sum_{c \in IV} Adm_c^{A-2}}{\sum_{c \in IV} Adm_c^{A-1}} \text{ et } \Theta_2 = R_{A-1}^{A-2} \times \frac{\sum_{c \in IV} Adm_c^{A-2}}{\sum_{c \in IV} Adm_c^{A-1}}.$$

Il est à peu près certain que ces deux séries, extrapolarions et répopulations, ne coïncideront pas. Toutefois, il est souhaitable de publier une et une seule série d'estimations pour toute zone pour tout moment. Il apparaît naturel de produire une série « composite » dont les extrêmes soient ancrés aux valeurs du recensement. La combinaison linéaire suivante peut jouer ce rôle en donnant plus d'importance à la collecte la plus récente :

$$R_{D,IV}^{A-2} = 0,2 \times \Theta_1 + 0,8 \times \Theta_2.$$

De même, en donnant les définitions appropriées à Θ_1 et Θ_2 , on ferait pour une commune E du groupe V :

$$R_{E,V}^{A-2} = 0,4 \times \Theta_1 + 0,6 \times \Theta_2.$$

Ces facteurs d'ajustement Θ devront être calculés pour des strates relativement fines de la population, des classes d'âge-sexe par exemple, de façon à préserver la plus grande souplesse démographique et géographique à l'ajustement des recensements. La qualité des fichiers administratifs et les disparités locales dicteront le niveau auquel l'ajustement peut être réalisé convenablement (des départements, des régions métropolitaines,...). On peut tenir un raisonnement analogue en grande commune si on remplace une « petite et moyenne commune » par une « adresse ».

Finalement, quand toutes les communes de tous les groupes auront été imputées, il est improbable que l'estimation du total d'une variable d'intérêt obtenu du fichier imputé (estimations détaillées) ne corresponde plus au total

La série des estimations des populations légales sont la troisième série tirée du recensement. Il s'agit des chiffres de population auxquels se réfèrent les textes de loi pour le financement des communes, le découpage électoral, la composition du conseil municipal, ...

La « population totale » légale d'une commune comprend :

- les individus dont la résidence principale est située dans la commune,
- ceux qui résident dans un établissement ou un logement collectif situé sur le territoire de la commune,
- ceux qui résident dans une autre commune mais qui ont gardé un logement dans leur commune d'origine, les personnes qui ont une résidence dans la commune et qui vivent dans un logement collectif d'une autre commune pour leur travail ou dans une autre commune pour leurs études,
- et les populations administrativement rattachées à la commune (forains, mariners, etc.).

On voit donc que ces populations ne peuvent être estimées qu'une fois l'ensemble du territoire couvert, c'est-à-dire qu'au moment des estimations détaillées.

3.4 Estimation de la variance d'échantillonnage

Il est prévu d'accompagner les estimations, globales et détaillées, d'une mesure de leur qualité statistique. Les travaux portant sur cet aspect ont débuté à l'automne 2001; l'option privilégiée pour l'instant est le recours à des tables de référence, comme on le fait pour l'enquête canadienne sur la population active, par exemple. Les variances d'échantillonnage seront vraisemblablement obtenues par

estime à partir des seules observations (estimations globales publiées deux ans auparavant). Il est donc convenu que les estimations détaillées soient calées aux estimations globales. Le niveau de calage dépendra encore une fois des tendances locales et de la qualité des estimations globales.

3.3 Estimations de populations légales

	A-6		A-5		A-4		A-3		A-2		A-1		A
Gr I	Adm	Adm	Adm	R	Adm	Adm	Adm	Adm	?	Adm	Adm	Adm	
Gr II	Adm	Adm	Adm	Adm	Adm	Adm	Adm	R_{A-3}^{A-2}	R_{A-2}^{A-2}	Adm	Adm	Adm	
Gr III	Adm	Adm	Adm	Adm	Adm	Adm	Adm	R_{A-2}^{A-2}	Adm	Adm	Adm	Adm	
Gr IV	R	Adm	Adm	Adm	Adm	Adm	?	Adm	R	Adm	Adm	Adm	
Gr V	Adm	Adm	Adm	Adm	Adm	Adm	?	Adm	Adm	Adm	R	Adm	
Total	SR Σ Adm		SR Σ Adm		SR Σ Adm		SR Σ Adm		SR Σ Adm		SR Σ Adm		Σ Adm

3.1 Estimations globales

A ce jour, les plans de diffusion prévoient la publication, au 31 décembre A, des résultats nationaux et régionaux pour l'enquête tenue en début d'année A; ces estimations forment les estimations globales pour l'année A. Il est aussi prévu de publier à la même date les résultats pour chacune des « petites et moyennes communes » visitées durant la campagne de collecte de l'année A.

3.2 Estimations détaillées

Les fichiers administratifs fourniront des informations complémentaires à un niveau de détail assez fin. Il sera alors possible de mesurer la distorsion entre ce qui a été observé et ce qui est inscrit au fichier pour des objets similaires (immeubles, îlots,...). Cette distorsion sur des agrégats bien déterminés peut se traduire en facteur de correction à appliquer aux données administratives de sorte que la somme corrigée de celles-ci corresponde bien aux estimations censitaires.

A ce jour, il est prévu d'exploiter les fichiers administratifs à un niveau d'aggrégation géographique (immeuble, îlot, district d'agent recenseur,...) qui renseigne sur les individus (âge, sexe d'après les fichiers de l'assurance maladie) ou leurs logements (fichiers de l'habitation, ci-dessous TH).

Les résultats détaillés relatifs à l'année A-2 seront mis à disposition au 31 décembre A (Il est prévu que l'acquisition et le traitement des fichiers administratifs prendront environ 2 ans); ces résultats détaillés seront l'amalgame d'observations faites par sondage (dans les grandes communes) ou recensement (dans les petites et moyennes communes) et de données synthétiques (dans les petites communes) et de données synthétiques.

Les données synthétiques seront obtenues à partir de la relation entre données observées et administratives sur un même point en un même instant. Par exemple, pour une commune C du Groupe II recensée en A-3, dont le recensement est établi à $R_{C,II}^{A-3}$, on obtient une imputation de son recensement pour l'année cible A-2 en faisant :

$$R_{C,II}^{A-2} = R_{C,II}^{A-3} \times \frac{Adm_{A-2}^{II}}{Adm_{A-3}^{II}} = R_{C,II}^{A-3} \times \frac{\sum_{c \in II} Adm_{A-2}^c}{\sum_{c \in II} Adm_{A-3}^c},$$

où Adm_c^c est la valeur des sources administratives pour la commune c et l'année a.

En régime permanent, pour une commune enquêtée en A-5 et A (voir la figure ci-dessous), on aura mesuré des variables sur les personnes (âge, sexe, activité, profession,...) et sur les logements (taille du ménage, nombre de pièces, mode d'occupation, confort,...) aux deux moments.

Le RLL sera mis à jour en continu à partir de permis de construire, de permis de démolir, de fichiers d'abonnés (eau, gaz, électricité, etc.), de renseignements fournis par les administrations locales et par l'observation directe sur le terrain. Ainsi, le RLL peut servir à la constitution d'une base de sondage « immeubles » en « grande commune ».

Dans chaque IRIS2000 (« îlots regroupés selon des indicateurs statistiques » zone homogène d'environ 2 000 habitants) de chaque « grande commune », on créera 5 groupes de rotation d'adresses sur le modèle du sondage des « petites et moyennes communes ». Trois strates supplé-

mentaires seront prévues dans chaque commune : une pour les immeubles d'activité (usines, entrepôts,...), une seconde pour les logements collectifs (établissements, collectivités, communautés, internats,...), et une dernière pour les adresses neuves.

On visitera chaque année un cinquième des immeubles d'activité pour s'assurer qu'ils sont toujours vides de logements (logement de gardien, ou espace converti à l'habitation); les logements éventuellement trouvés dans de tels immeubles seraient considérés autoréprésentatifs parce qu'exceptionnels. L'ensemble des logements collectifs sera couvert chaque année; un cinquième d'entre eux seront visités alors que l'effectif des quatre autres cinquièmes sera éventuellement mis à jour par enquête téléphonique. Finalement, les immeubles d'habitation neufs seront insérés dans les groupes de rotation.

Comme décrit plus haut, les groupes de rotation d'adresses seront visités à tour de rôle au cours d'une période de 5 ans. Un sous-échantillon d'adresses correspondant à 40 % des logements du groupe sera alors tiré. Dans chaque adresse sélectionnée, l'ensemble des logements sera enquêté.

En résumé, l'échantillon annuel comptera environ 8 millions de bulletons individuels, 6 millions des « petites et moyennes communes » et 2 millions des « grandes communes ».

3. ESTIMATIONS GLOBALES ET DÉTAILLÉES

En régime courant, trois séries d'estimations seront produites et diffusées chaque année : une série d'estimations des populations légales, une série d'estimations détaillées – dont on tirera les populations légales – et une série d'estimations globales servant au calage des précédentes.

Pour chaque groupe de rotation, on voit et les quartiles et l'étendue de la distribution; il est intéressant de noter la superposition des diagrammes. La variable « Nombre de femmes âgées de 20 à 39 ans » a été utilisée pour la composition des groupes; le nombre de résidences principales, ni aucune des variables associées au ménage ou au logement, n'intervient pas dans l'établissement de l'équilibre.

NOMBRE DE FEMMES DE 20 À 39 ANS

Rhône-Alpes, 1990

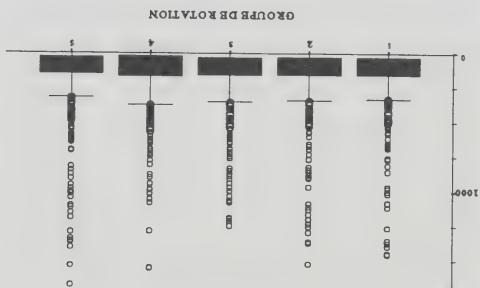


Figure 1.

NOMBRE DE RÉSIDENCES PRINCIPALES

Rhône-Alpes, 1990

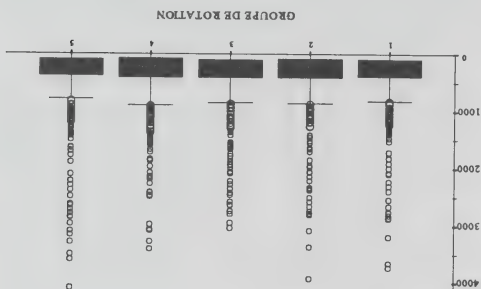


Figure 2.

Chaque année, on fera le recensement (exhaustif) de la population et des logements de toutes les communes d'un des groupes de rotation. Ainsi, chacune des « petites » communes communes sera recensée une fois tous les 5 ans, et toutes les « petites » communes communes à raison d'un cinquième par année.

2.2 Les grandes communes

Le sondage en « grande commune » utilisera le « répertoire d'immuables localisés » (RIL). Ce répertoire est une

Alfin de définir précisément le cadre d'organisation entre communes ainsi que les processus de validation des différentes étapes.

Maîtrise des coûts : la réalisation de la collecte sur un cycle de 5 ans permettra d'étaler les charges financières affectées à l'opération. En ce qui concerne les communes de plus de 10 000 habitants, la charge du recensement

de population en régime courant. En revanche, la charge de population en régime permanent sera équivalente au coût d'un

recensement général, ce qui permet de réaliser la réforme sans surcoût budgétaire. Cependant, les premières années de collecte pourront supporter un budget quelque peu

supérieur à ce montant afin de permettre le rodage du processus.

2. STRATÉGIE D'ÉCHANTILLONNAGE

La commune constitue le point d'ancrage de la rénovation : les « petites et moyennes communes » (celles de moins de 10 000 habitants) seront sondées au taux (moyen) d'un cinquième par an et tous leurs logements seront

visités; toutes les « grandes communes » seront visitées chaque année, mais seulement une fraction de leurs

2.1 Les petites et moyennes communes

Considérons d'abord le domaine des « petites et moyennes communes ». Dans chaque région, cinq groupes de rotation seront créés à partir des renseignements du recensement de la population (RP99) par tirage d'échantillons équilibrés (Deville, Tillie (1999, 2000)) sur la distribution

âge-sexe des communes; cette approche devrait permettre de minimiser les variations interannuelles dues au seul

sondage. Les figures 1 et 2 illustrent comment les 5 groupes de rotation sont équilibrés. Ces deux figures donnent les

« diagrammes à moustaches » de deux variables mesurées sur les 2 811 petites et moyennes communes de Rhône-Alpes au Recensement général de la population de 1990.

La rénovation du recensement français

J.-M. DURR et J. DUMAIS¹

RÉSUMÉ

Il devient de plus en plus difficile de mener des recensements de façon traditionnelle. Quand on a la possibilité d'interconnecter des fichiers administratifs, on ouvre une alternative intéressante à la pratique de recensements périodiques (Laihoen 2000; Borchsenius 2000). On retrouve ce type de proposition dans un article récent de Mathan (2001). La rénovation développée à l'INSEE repose sur le concept de « recensement continu » dont l'idée remonte à Kish (1981, 1990) et Horvitz (1986). Une première approche envisageable en France peut être trouvée dans Deville et Jacod (1996). Le présent article fait le point des développements méthodologiques depuis que l'INSEE a mis en route son Programme de rénovation du recensement de la population.

MOTS CLÉS : Échantillonnage équilibré; recensement; recensement continu; calage.

1. INTRODUCTION

1.1 Les raisons de la rénovation

La France conduit depuis de nombreuses années des recensements afin de déterminer la population légale de chaque commune de ses circonscriptions administratives et de caractériser les niveaux géographiques, des quartiers des communes au pays dans son ensemble. Ainsi, le recensement de 1999 s'est déroulé selon le schéma habituel : dépôt et reprise des questionnaires par des agents recenseurs, organisation, assistance technique et contrôle par l'INSEE, exécution par le Maire en tant que représentant de l'État. Toutefois, certains éléments nous ont conduits à revoir ce dispositif.

Tout d'abord, l'intervalle intercensitaire a tendance à s'allonger. En effet, la périodicité des recensements n'est pas inscrite dans la loi, et la date de chaque recensement est fixée par décret. On est passé de recensements quinquennaux avant la guerre, à des écarts entre les recensements de 7, puis 8 ans. Le dernier recensement, prévu initialement en 1997, a été repoussé à 1999, soit 9 ans après le précédent, pour des raisons budgétaires. De plus, le public ne comprend pas toujours la nécessité d'une opération aussi lourde dans un contexte de fichiers administratifs toujours plus nombreux, même s'il redoute par ailleurs leur utilisation croisée. Enfin, le mouvement de décentralisation que connaît la France depuis plus de vingt ans a généré de nombreux besoins de données statistiques afin d'éclairer les politiques locales. Le recensement, source d'information locale par excellence, doit donc s'adapter et fournir des données plus fraîches et toujours finement localisées.

C'est pourquoi un programme de rénovation du recensement de la population a été engagé à l'Insee dès la

fin des années 90. La France ne disposant pas de registre de population et le contexte national ne s'y prêtant pas, il a donc été décidé d'envisager une voie intermédiaire combinant la réalisation d'enquêtes annuelles par sondage et l'utilisation de fichiers administratifs non nominatifs que l'Insee est habilitée à utiliser à des fins exclusivement statistiques. Pour les communes dont la population est inférieure à un seuil, fixé pour l'instant à 10 000 habitants, les enquêtes seront exhaustives et auront lieu chaque année par roulement au cours d'une période de cinq ans. Pour les autres communes, une enquête par sondage sera effectuée chaque année, la totalité du territoire de ces communes étant prise en compte au terme de la même période de cinq ans. Pour mener à bien cette rénovation, un cadre juridique nouveau s'est avéré nécessaire. Le Conseil d'État, consulté sur le projet, a préconisé, dans son avis du 2 juillet 1998, que le gouvernement soumette au Parlement un projet de loi.

Outre la nécessité de renforcer l'assise légale du recensement, il a considéré que le changement important des modalités d'élaboration des chiffres de population, alors que plus de 200 textes législatifs ou réglementaires s'y réfèrent, nécessitait de passer par la voie législative. Dans le cadre ainsi défini, le projet de loi vise essentiellement à définir les principes et à fixer les règles de base applicables à l'organisation du recensement. L'opération est placée sous la responsabilité et le contrôle de l'État : l'Insee organise le cadre de la collecte (concepts, protocoles), réalise le tirage des échantillons, veille à la qualité des informations collectées, exploite les données et les diffuse. Les communes, en tant que collectivités locales, préparent et réalisent les enquêtes de recensement. En compensation, l'État leur verse une dotation financière. Ces dispositions clarifient le rôle de chacun des partenaires et les responsabilisent quant à leur incombent.

¹ Jean-Michel Durr, Programme de rénovation du recensement de la population, INSEE, Direction générale, 18 boul. Adolphe Pinard, 75675 Paris CEDEX 14, France; Jean Dumais, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa K1A 0T6, Canada; Cet article a été préparé pendant que l'auteur était en contrat au Programme de rénovation du recensement de la population, INSEE.

REMERCIEMENTS

Le présent article présente les résultats de travaux de recherche et d'analyse entrepris par les employés du Census Bureau. Il a fait l'objet, de la part du Bureau, d'un examen de portée plus limitée que celui auxquelles sont soumis les publications officielles. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et d'en favoriser la discussion.

BIBLIOGRAPHIE

- ALEXANDER, C.H. (1993). A continuous measurement alternative for the U.S. Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 486-491.
- ALEXANDER, C.H. (1997). Questions relatives au plan de sondage des I' « American Community Survey » et résultats préliminaires des *Recueil du Symposium 97, Nouvelles Orientations pour les Enquêtes et les Recensements*, 211-217.
- ALEXANDER, C.H. (1998). Recent developments in the american community survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 92-100.
- ALEXANDER, C.H., et WETROGAN, S. (2000). Integrating the american community survey and the intercensal demographic estimates program. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 295-300.
- BOUNPANE, P. (1986). How increased automation will improve the 1990 census. *Journal of Official Statistics*, 4, 545-553.
- BUTANI, S., ALEXANDER, C. et ESPOSITO, J. (1999). Using the american community survey to enhance the current population survey: Opportunities and issues. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Statistical Policy Working Paper 29*, 3, 102-111.
- COCHRAN, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (B)*, 4, 102-118.
- COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- ECKLER, A.R. (1972). *The Bureau of the Census*. New York: Praeger Publishers.
- HANSEN, M.H., MADOW, W.G. et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 384, 776-793.
- HAUSER, P.M. (1942). Proposed annual census of the population. *Journal of the American Statistical Association*, 37, 81-88.
- HERRIOT, R.A., BATEMAN, D.B. et MCCARTHY, W. F. (1989). The Decade Census Program - A new approach for meeting the nation's needs for sub-national data. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-355.
- KISH, L., LOVEJOY, W. et RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.
- KISH, L. (1979b). Rolling samples instead of censuses. *Asian and Pacific Census Forum*, (G1), August 1979, 1-2, 12-13.
- KISH, L. (1981). *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*. Washington, D.C., U.S. Government Printing Office.
- KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright). New York: Academic, 72-84.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 1-12.
- KISH, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons.
- KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-86.
- KISH, L. (1997). Periodic and rolling samples and censuses. *Chapitre 7 dans Statistics and Public Policy*, (Ed. Bruce D. Spencer). Clarendon Press, Oxford.
- KISH, L. (1998). Spacetime variations and rolling samples. *Journal of Official Statistics*, 14, 1, 1998, 31-46.
- KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 2, 147-158.
- KISH, L. (2001). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, à paraitre dans 2001.
- KISH, L., et VERMA, V. (1983). Census plus samples: Combined uses and designs. *Bulletin of the International Statistical Institute*, 50(1), 66-82.
- KISH, L. et VERMA, V. (1986). Complete Censuses and Samples. *Journal of Official Statistics*, 2, 381-93.
- MELNICK, D. (1991). The census of 2000 A. D. and beyond. *Reviews of Major Alternatives for the Census in the Year 2000*. U.S. Government Printing Office, Washington, D.C., August 1, 1991, 60-74.
- MOONEY, H.W. (1956). Methodology in two California Health Surveys. *U.S. Public Health Monograph*, 70.
- NATIONAL ACADEMY OF SCIENCES (1994). *Country People in the Information Age*. Duane L. Steffey et Norman M. Bradburn, eds. National Academy Press, Washington, D.C.
- NATIONAL ACADEMY OF SCIENCES (1995). *Modernizing the U.S. Census*. Barry Edmonston et Charles Schultze, eds. National Academy Press, Washington, D.C.
- PURCELL, N.J., et KISH, L. (1979). Estimation for Small Domains. *Biometrics*, 35, 365-384.
- SAWYER, T. C. (1993). Rethinking the Census: Reconciling the Demands for Accuracy and Precision in the 21st Century. *Présente au Research Conference on Undercounted Ethnic Populations*, 7 Mai, 1993.
- TUPEK, A. R., WAITE, P. J. et CAHOON, L. S. (1990). Sample Expansion Plans for the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 72-77.

questionnaire détaillé du recensement et la CPS par une seule enquête. Contrairement à d'autres questions sur lesquelles nous avions des discussions animées, Leslie avait adopté sur ce point une attitude de « non-intervention ». À mon avis, il considérerait cette décision comme une question de compromis concernant la qualité que les organismes gouvernementaux devaient résoudre eux-mêmes. Les deux raisons principales de notre décision sont les suivantes.

Nous ne pouvons évaluer adéquatement le taux mensuel de chômage au moyen d'une enquête postale. La mesure correcte du taux de chômage nécessite des questions complexes que l'on ne pourrait poser au moyen d'un instrument envoyé par la poste, comme celles visant à s'assurer qu'une personne qui « recherche du travail » a activement entrepris des démarches pour trouver un emploi. (Voir Butani, Alexander et Esposito 1999). L'enquête supplémentaire du Recensement de 2000, sur-estimation importante du taux national de chômage en 2000 (5,3 % contre 4,0 % pour la CPS). Un écart semblable a été observé pour le Recensement de 1990.

Comparativement à la CPS, les taux mensuels calculés d'après les données des enquêtes postales seraient produits avec un retard important. En outre, le fait qu'il soit impossible de réaliser toutes les interviews par la poste pour un panel donné dans le mois désigné biaise les estimations mensuelles (voir la section 5.5 plus haut). La production de moyennes mobiles trimestrielles au lieu d'estimations mensuelles, solution fréquemment suggérée par Leslie (par exemple, Kish 1999), réduirait ces problèmes, mais le rapport mensuel sur le chômage est un indicateur économique indispensable aux États-Unis.

Parce qu'elle est conçue si rigoureusement pour remplacer un questionnaire détaillé, l'ACS n'illustre pas pleinement la souplesse envisagée par Leslie pour une enquête à échantillons successifs. Dans d'autres circonstances, pour une population plus petite et un besoin moins grand de l'« enquête à questionnaire détaillé » pour produire des estimations pour des domaines très petits, ou dans le cas

d'exigences moins rigides quant aux délais de production et aux questions à poser, il sera peut-être possible qu'une enquête sur la population active à échantillons successifs permette de répondre à la demande de données régionales. Grâce à l'ajout d'un panel double ou d'autres composantes (Kish 1998, pages 40 et 41), on pourrait atteindre une gamme encore plus grande d'objectifs.

9. CONTRIBUTION : CONCEPTUELLE, PERSONNELLE ET PRATIQUE

La longue liste d'articles publiés par Leslie Kish au sujet des échantillons successifs démontre clairement l'intensité et la ténacité de sa campagne en faveur de ce qu'il considérait comme une idée importante. L'évolution de cette idée au fil de ses articles illustre aussi la grande attention qu'il accordait aux questions « conceptuelles » concernant les objectifs fondamentaux de qualité d'une enquête. Qu'essayons-nous de faire? Quel est le rapport entre le choix du plan d'enquête et ce que nous essayons de réaliser, et pourquoi? Cette forme d'orientation, qui est essentielle lors du lancement d'un programme d'enquête quand il faut résoudre les « grandes questions », permet de faire la distinction entre les idées qui sont abandonnées rapidement et celles qui font leur chemin.

L'appui personnel que Leslie offrait aux autres statisticiens dépassait de loin le cadre de ses articles. Alors que je n'étais nullement l'un de ses collègues les plus proches, il me prodiguait régulièrement des conseils personnels ou des encouragements lorsqu'il le jugeait nécessaire. L'expression « *still rolling* » du titre de la version anglaise de cet article était celui que je donnais aux messages électroniques que je lui envoyais pour le tenir au courant de la situation pénibleuse de l'ACS aux diverses étapes du cycle budgétaire annuel, c'est-à-dire la plupart du temps. Il répondait brièvement par courriel, mais transmettait toujours les messages importants sous forme de lettres manuscrites.

Enfin, si l'on s'en tient à ses articles, il est évident que Leslie a toujours été une personne pratique, même dans ses moments de réflexion les plus théoriques, et que ses articles ne peuvent être appréciés comme il se doit si l'on ignore ce qui se passait dans le monde des enquêtes à l'époque où il les a rédigés. Lorsque je réfléchis à ses articles sur les échantillons successifs, je décèle de nombreux commentaires, au sujet de détails aussi bien que de principes généraux, qui visaient à éclaircir certaines décisions que le Censur Bureau devait prendre à l'époque. J'imagine que l'ensemble de ses travaux contient divers messages destinés spécifiquement à aider quelqu'un, quelque part dans le monde, qui devait prendre à l'époque une décision pratique concernant un plan de sondage.

nous ne leur accorderons pas autant d'importance dans nos publications ou nos sites Web.

Ces produits de données de l'ACS dont nous prévoyons la publication visent à inciter les analystes à utiliser la même période de cumul lorsqu'ils comparent des régions de tailles différentes, car ne pas le faire serait perçu comme injuste pour les régions plus petites. Ce faisant, nous avons accepté la notion de « cumul asymétrique » en ce qui concerne les niveaux de détail géographique, mais pas nécessairement à l'intérieur d'un niveau de détail géographique particulier. Par exemple, nous pourrions choisir une période d'une année pour comparer les États, mais nous recommanderions une période de cinq années pour tous les comtés repris dans un tableau comparant les grands et les petits comtés. Cette dernière recommandation nous écarte quelque peu de la démarche de Kish (1998, pages 42-43) qui permettrait d'utiliser des tableaux contenant des estimations sur un an pour les grands comtés, et des moyennes sur trois ans pour ceux de taille moyenne, et des moyennes sur cinq ans pour ceux de petite taille. Il sera intéressant de voir quelle pratique les utilisateurs des données adopteront à cet égard.

7. PONDERATION DES ANNÉES LORS DU CUMUL PLURIANNUEL

Kish (1998) fait remarquer qu'il existe plusieurs options pour la pondération des cumuls pluriannuels. S'il existe 10 moyennes annuelles \bar{y}_i , alors il existe de nombreux choix de $\bar{y} = \sum w_i \bar{y}_i$, où $\sum w_i = 1$, pour les cumuls sur 10 ans.

Pour les cumuls de données de l'ACS sur cinq ans et d'autres cumuls pluriannuels, discutés à la section 6, nous prévoyons accorder un poids égal à toutes les années de référence dans les produits de données publiés types, c'est-à-dire $w_i = 0,2$ pour la moyenne sur cinq ans. Il s'agit là d'un sujet de désaccord avec Kish (1998) qui nous priait de considérer d'autres solutions, particulièrement des poids de la forme $w_{i+1} = Cw_i$, avec $C > 1$.

Lorsque l'on pense à des poids inégaux, l'une des questions sous-jacentes est de savoir quel problème statistique nous essayons de résoudre. Si nous prenons le cumul des données de 2003 à 2007 comme exemple, l'objectif est-il de fournir :

- une estimation « directe fondée sur le plan de sondage » de la moyenne historique pour la période de 2003 à 2007;
- une estimation « basée sur un modèle » pour la valeur de 2007;
- une estimation « directe fondée sur le plan de sondage » de la moyenne historique pour la période de 2003 à 2007;

8. NE PAS COMBINER LA CPS ET L'ACS

« D'importantes questions restent à débattre et à étudier. Peut-être à jamais, le sujet étant devenu une « branche d'activité en croissance. » [Traduction]

(1998, page 40) aura le dernier mot à ce sujet :

Cependant, après avoir relu l'article de Kish (1998), j'interprète maintenant sa vision de la moyenne pondérée comme correspondant à la troisième formulation, c'est-à-dire un estimateur fondé sur le plan de sondage d'un paramètre de population plus à jour. Cette solution permet d'éviter les questions quant à la validité du modèle aux applications générales. Une question demeure : comment justifier et réaliser une solution consensuelle? En outre, l'application de poids inégaux a tendance à augmenter les erreurs-types des moyennes pluriannuelles. Toutefois, Kish (1998, page 40) observe des variations durant la période de cinq ans.

Lorsque l'on observe des variations durant la période de cinq ans, j'interprète maintenant sa vision de la moyenne pondérée comme correspondant à la troisième formulation, c'est-à-dire un estimateur fondé sur le plan de sondage d'un paramètre de population plus à jour. Cette solution permet d'éviter les questions quant à la validité du modèle aux applications générales. Une question demeure : comment justifier et réaliser une solution consensuelle? En outre, l'application de poids inégaux a tendance à augmenter les erreurs-types des moyennes pluriannuelles. Toutefois, Kish (1998, page 40) observe des variations durant la période de cinq ans.

Précédemment, j'ai considéré la décision comme étant entre tableaux. Maintenir l'additivité dans les tableaux et la comparabilité pour toutes les régions et toutes les caractéristiques afin de 2007 avec une prévision pour 2007 fondée sur les années 2003, ..., 2006 à condition que la même formule soit utilisée comme étant la combinaison d'une estimation directe pour la région en question. Le problème peut être considéré ou des hypothèses au sujet de la série chronologique pour comme étant une estimation pour 2007 nécessite un modèle

Leslie a souvent déclaré qu'il était heureux de voir son idée mise en œuvre dans l'ACS, mais je pense qu'il était déçu que nous n'ayons pas essayé de remplacer le

5.5 Date de référence des questionnaires, étant donné une période étendue d'interviews

Pour chaque panel mensuel de l'ACS, les interviews ont lieu sur une période de trois mois, deux mois étant accordés pour le retour du questionnaire dûment rempli par la poste et le suivi téléphonique, avant le lancement des visites sur place plus coûteuses le troisième mois. Donc, les données recueillies pour juin, par exemple, englobent les réponses rapides par la poste des membres du panel de juin, les réponses tardives par la poste et les interviews téléphoniques pour le panel de mai, ainsi que les cas de suivi sur place pour le panel d'avril. Cette situation oblige à se demander s'il faut que les répondants fournissent les réponses qui correspondent à la date d'envoi par la poste, c'est-à-dire la meilleure option en ce qui concerne le biais d'échantillonnage, ou celles qui correspondent au moment où les questions sont posées, c'est-à-dire la meilleure option en ce qui concerne l'erreur de réponse et d'autres erreurs non dues à l'échantillonnage, particulièrement pour les personnes qui ont changé d'adresse.

Compte tenu de ces compromis concernant la qualité des données, nous avons choisi d'utiliser comme date de référence la « date courante » et de recueillir les données sur les caractéristiques des membres du ménage au moment de l'interview. Nous avons pris cette décision notamment parce que nous estimons que les erreurs non dues à l'échantillonnage seront plus difficiles à évaluer que le biais d'échantillonnage. En outre, les biais d'échantillonnage qui entraînent les estimations mensuelles auront tendance à se compenser au cours de l'année. Il s'agit de l'une des raisons de limiter l'ACS à la production d'estimations annuelles et pluriannuelles.

5.6 Utilisation d'estimations intercensitaires de population pour le contrôle des poids de sondage

Le Census Bureau a mis en place un programme d'estimations démographiques « intercensitaires » (Leslie qualifierait ces estimations de « postcensitaires », réservant le terme « intercensitaires » pour les estimations réalisées entre deux recensements achevés.) Fondé sur des modèles démographiques, ces modèles permettent de mettre à jour les données du recensement antérieur, d'après les données des dossiers de l'état civil et d'autres dossiers administratifs. Ces estimations servent de contrôles indépendants de la pondération, ou de « facteurs de stratification a posteriori », pour la plupart des enquêtes nationales (voir Kish 1965, pages 90-92). L'ajustement des poids de sondage afin qu'ils concordent avec les valeurs de contrôle permet de réduire la variance des estimations d'enquête, de rajuster les données pour tenir compte des différences de couverture selon l'âge, le sexe, la race ou l'origine hispanique, et d'augmenter la cohérence d'une enquête à l'autre. De façon analogue, dans le cas du questionnaire détaillé du recensement, on utilise les dénombremen-

recensement comme contrôles de la pondération.

6. CUMULS DIFFÉRENTS SELON L'OBJECTIF

Habituellement, aucun contrôle de la pondération n'existait pour les domaines géographiques les plus petits, du moins pas au niveau de détail démographique existant pour les régions plus grandes. La production prévue de totaux de contrôle plus détaillés pour la pondération des données de l'ACS est décrite dans Alexander et Wrogan (2000). Certaines améliorations résulteront de l'utilisation de meilleures sources de données administratives, mais l'ACS proprement dite fournira des données sur les variations de population qui pourront être intégrées dans les modèles démographiques. La différence entre la « règle du résident courant » appliquée pour l'ACS et la « règle du résident usuel » appliquée pour le recensement complique le problème; l'ACS comprend une question au sujet des résidents ayant occupé le logement une partie de l'année afin de permettre la correction pour cette différence de concept. Afin de faciliter cette intégration des données d'enquête dans les modèles démographiques, particulièrement en vue d'élaborer des mesures de l'erreur pour les estimations résultantes, le Census Bureau essaie de mettre au point des versions « statistiques » des modèles démographiques utilisés pour produire les estimations intercensitaires de population. Les efforts en vue d'intégrer les approches statistiques et démographiques s'inspirent de Purcell et Kish (1979).

Afin d'atteindre l'objectif principal de l'ACS, c'est-à-dire remplacer le questionnaire détaillé du recensement en tant que source de statistiques descriptives détaillées, nous obtenir un produit de données comparable au « fichier sommaire » traditionnellement fondé sur le questionnaire détaillé. Cette période est la plus courte pour laquelle l'erreur d'échantillonnage de l'ACS est jugée raisonnablement proche de celle relative au questionnaire détaillé du recensement. Le fichier de données sur cinq ans engloberait toutes les tailles et toutes les catégories de zones géographiques. En ce qui concerne l'affectation des fonds gouvernementaux basée sur l'évaluation des besoins courants, les études en simulation donnent à penser que le cumul des données sur trois ans serait préférable au cumul sur cinq ans, ce qui revient à sacrifier la précision pour une plus grande actualité (Alexander 1998).

Pour les régions individuelles, les données les plus importantes publiées seront les moyennes sur un an pour les régions comptant plus de 65 000 habitants et les moyennes sur trois ans pour celles comptant plus de 20 000 habitants, en plus des moyennes sur cinq ans produites pour toutes les régions dont la population est inférieure à ces seuils seront offertes pour des utilisations plus « complexes » dans des modèles de séries chronologiques et pour indiquer les variations importantes des moyennes pluriannuelles, mais

soi, l'élimination du questionnaire détaillé, sans autres améliorations fondamentales, permette une économie suffisante pour financer l'ACS.

5. CERTAINES VARIANTES DU PLAN DE SONDAGE DE BASE ET CERTAINS PROBLÈMES

5.1 Échantillons en grappes à plusieurs degrés

d'échantillonnage, égal au taux le plus élevé nécessaire initial est sous-échantillonné pour obtenir le taux d'échantillonnage souhaité pour chaque strate pour l'année de référence. La sélection de sous-échantillons, qui permet d'éviter le chevauchement avec la totalité des superéchantillons antérieurs, oblige uniquement à suivre le taux d'échantillonnage pour la première étape.

5.3 Mises à jour de la base de sondage

En pratique, la population varie légèrement pour chaque panel. De nouvelles adresses sont ajoutées à la base de sondage. Certaines anciennes adresses n'existent plus et peuvent être supprimées de la liste, ou y être maintenues et supprimées uniquement après que l'on ait essayé de communiquer avec les personnes vivant à ces adresses. Ceci ne présente aucun problème conceptuel fondamental et signifie qu'un « recensement continu » ne consiste pas nécessairement à communiquer avec toute unité de population ayant déjà existé, puisque certaines unités peuvent disparaître trop rapidement après leur création pour tomber dans l'échantillon.

Pour éviter l'enregistrement des divers taux d'échantillonnage conditionnels appliqués pour différentes « cohortes » d'adresses ajoutées au moment des mises à jour du fichier maître d'adresses exécutées à diverses périodes, nous avons jugé pratique d'attribuer des « échantillons rétrospectifs » artificiels en sélectionnant des adresses à partir de chaque ensemble de nouvelles adresses, non seulement pour le panel courant, mais aussi pour les anciens panels. Ces unités ne sont pas interviewées, puisque la période d'existence du panel auquel elles sont assignées est révolue, mais elles peuvent ainsi être évitées durant la sélection sans remise des futurs panels.

5.4 Que se passe-t-il après le panel k ?

Une question qui, autant que je le sache, n'a pas été abordée explicitement par Leslie est la façon de tirer l'échantillon pour le panel $k + 1$. Je crois qu'il supposait que le panel $k + 1$ serait le même que le panel 1, que le panel $k + 2$ serait une répétition du panel 2, et ainsi de suite. Cette démarche donne de bons résultats pour un échantillon aléatoire simple, mais ne fonctionne pas aussi bien pour un échantillon systématique destiné à étaler l'échantillon sur une liste triée géographiquement, car, comme la base de sondage évolue avec le temps, le panel 1 ne garde pas son espace uniformément.

Nous prévoyons sélectionner le panel $k + 1$ et les panels suivants sous forme d'échantillons systématiques frats. Chacun sera sélectionné de sorte qu'il ne chevauche aucun des $k - 1$ panels antérieurs, si bien que nous aurons systématiquement k panels consécutifs non chevauchants, sans devoir nous inquiéter du chevauchement avec les panels sélectionnés avant cela.

L'ACS est réalisée auprès d'un échantillon systématique à un degré non mis en grappe, parce que les objectifs comprennent la production annuelle de données pour tous les petits domaines géographiques, comme les secteurs de recensement ou les groupes d'îlots. Les discussions publiées par Kish (1981, 1998) montrent clairement que la méthode des échantillons successifs s'applique aussi à l'échantillonnage par grappes et à l'échantillonnage à plusieurs degrés, ainsi qu'à diverses probabilités de sélection. Cependant, pour que l'on puisse parler d'« échantillons successifs », il faut que les unités primaires d'échantillonnage forment elles-mêmes des échantillons successifs. Un plan de sondage comportant un ensemble fixe d'unités primaires d'échantillonnage (UPB), avec échantillonnage successif dans chaque UPB, produit un « échantillon cumulatif représentatif » (Kish 1998).

Leslie insistait sur le fait que la proposition d'Herrnstein et coll. (1989) ne correspondait pas à ce qu'il entendait par « échantillons successifs ». Pourtant, la proposition semble correspondre à la définition si l'on considère les États comme des UPB. Selon moi, ce fait prouve qu'il est implicitement nécessaire qu'un plan de sondage à échantillons successifs produise un échantillon probabiliste représentatif utile lors de chaque période de référence, pour chaque domaine étudié, condition qu'il n'est pas vérifiée si les UPB sont des États. Autrement dit, les grappes, ou UPB, doivent être beaucoup plus petites que le plus petit domaine étudié. (Voir Kish 1998, page 38.)

5.2 Taux d'échantillonnage variable

Selon Kish (1998, section 4), un plan de sondage à échantillons successifs peut prévoir l'application d'un taux d'échantillonnage variable selon la strate. Ce genre de plan peut être compliqué, particulièrement si les taux d'échantillonnage varient avec le temps, car la probabilité conditionnelle de sélectionner une unité (sans remise) pour le j ème panel dans la h ème strate dépend des taux d'échantillonnage utilisés pour les panels antérieurs dans cette strate. La situation est encore plus compliquée si la strate varie au cours du temps, par exemple si les limites des unités gouvernementales changent.

Par souci de simplification, dans le cas de l'ACS, nous tirons l'échantillon en deux étapes. À la première, nous sélectionnons des « superéchantillons successifs » en appliquant pour chaque panel et chaque année un même taux

« mesure continue » pour remplacer le questionnaire détaillé du recensement a été envisagée durant les travaux de recherche pour le Recensement de 2000 qui ont débuté en 1992. Le plan de sondage à échantillons successifs de Kish a finalement été proposé parce qu'il offrait non seulement une certaine souplesse pour la production d'estimations, mais aussi la possibilité de recueillir efficacement des données (Alexander 1993, 1997; National Academy of Sciences 1994, 1995). De mémoire, les articles les plus influents ont été ceux de Kish (1981, 1990); ceux de Kish et Verma (1983, 1986) ont également été consultés. La « mesure continue » a été renommée plus tard « *American Community Survey (ACS)* ». L'ACS proposée n'a pas été adoptée pour le Recensement de 2000, mais, après des essais limités réalisés de 1996 à 1998, la méthodologie de l'ACS a été mise en œuvre dans 36 comtés pour la période de 1999 à 2001, de sorte que l'on puisse comparer les résultats de cette enquête aux données du questionnaire détaillé du Recensement de 2000. Un essai à grande échelle a également été réalisé en 2000, au moyen d'un échantillon annuel représentatif des États comportant environ 700 000 adresses. Cet essai, nommé *Census 2000 Supplementary Survey*, visait à recueillir les données du questionnaire détaillé indépendamment du recensement, à l'aide du questionnaire de l'ACS. La réalisation de l'enquête supplémentaire se poursuit en 2001 et en 2002 dans le cadre de la transition à l'ACS.

4. PLAN DE L'AMERICAN COMMUNITY SURVEY

L'ACS, qui débutera en 2003, est financée par le Congrès et porte sur un échantillon mensuel d'environ 250 000 adresses renouvelé entièrement au début de chaque mois, ce qui donne un plan à échantillons mensuels successifs pour lequel le taux moyen d'échantillonnage est d'environ $F = 480$ ou un échantillon annuel pour lequel $F = 40$. On utilisera pour l'enquête une valeur de $k = 60$, et la période de cumul d'échantillon la plus courte pour laquelle des données seront publiées sera l'année civile. L'ACS sera réalisée par la poste, avec suivi des non-répondants par téléphone. En outre, un échantillon aléatoire d'un tiers des non-répondants persistants sera sélectionné aux fins d'un suivi sur place. Pour les domaines ayant un taux de réponse moyen, avec une valeur mensuelle $F = 480$, l'erreur-type pour une moyenne estimée sur cinq ans d'après les données de l'ACS sera un peu plus importante que pour l'estimation correspondante calculée d'après les données du questionnaire détaillé du recensement — habituellement environ 1,33 fois plus importante. Ce résultat a été considéré « suffisamment proche » pour la plupart des applications, étant donné l'avantage quant à l'actualité des données et à la diminution prévue du taux de données manquantes, grâce à la création

d'un groupe permanent d'intervieweurs. Dans les régions où le taux de réponse par la poste est inférieur à la moyenne, le sous-échantillonnage aux fins du suivi des non-répondants réduira la taille effective de l'échantillon. Cette situation est due non seulement à la réduction du nombre d'interviews, mais aussi au fait que des poids de sondage inégaux entraînent habituellement un effet de plan de sondage plus important (Kish 1965, pages 429-431). Pour tenir compte de ceci, on réalisera vraisemblablement l'ACS en appliquant un taux plus élevé de sous-échantillonnage pour la non-réponse dans les régions à faible taux de réponse, équilibré par un taux plus faible d'échantillonnage dans les régions où le taux de réponse par la poste est supérieur à la moyenne. Les détails de cette procédure restent à déterminer. On procédera aussi à un suréchantillonnage des adresses dans le cas des petites unités gouvernementales, comme cela était le cas pour l'échantillon recevant le questionnaire détaillé du recensement. L'un des progrès importants de la dernière décennie, qui a rendu possible la réalisation de l'ACS (Kish (1981) propose comme méthode de rechange celle des « listes successives cumulatives », mais cette option rendrait onéreux le calcul d'estimations régulières pour tous les domaines les plus petits, comme les secteurs de recensement), est la mise en place du programme créé par le Census Bureau pour tenir à jour en permanence un fichier maître d'adresses (*Master Address File/MAF*) couplé à notre base de données géographiques TIGER. La source principale de mise à jour des adresses au cours de la décennie est le *Delivery Sequence File (DSF)* tenu à jour par le Service des postes. Le Census Bureau est en train de mettre en œuvre un programme de modernisation du MAF/TIGER qui améliorera les mises à jour d'après le DSF grâce à des fichiers de données transmis par les gouvernements locaux et d'autres sources administratives et grâce au ciblage de nouvelles adresses, dans des régions plus rurales, par des intervieweurs préposés à l'ACS et à d'autres enquêtes. En fait, pour produire les échantillons mensuels, on sélectionnera un échantillon annuel d'après le MAF du mois de septembre antérieur et on le divisera en 12 panels mensuels. En février, on sélectionnera d'après le DSF un échantillon supplémentaire de nouvelles unités que l'on répartira entre les mois restants de l'année.

Le remplacement du questionnaire détaillé du Recensement de 2010 par l'ACS est l'une des composantes du programme de remaniement du Recensement de 2010. Ce programme comprend aussi la modernisation du MAF/TIGER, ainsi qu'un programme de recherche et d'essais précoces en vue d'automatiser, de simplifier et d'améliorer les opérations du Recensement de 2010. En principe, le coût budgété de cette combinaison d'améliorations pour le cycle complet de 10 ans sera inférieur à celui du Recensement de 2000 en raison de la vision de 2010. Ce plan diffère considérablement de la vision de l'ACS décrite dans National Academy of Sciences (1994, chapitre 6; 1995, chapitre 6), où je formulais l'espoir qu'en

au fait qu'il permet de réaliser divers compromis entre les niveaux de détail spatial, temporel et démographique.

Leslie Kish a laissé à ses collègues la tâche de généraliser ses idées en une « théorie de la combinaison de populations » (Kish 1999, 2001). Lors de la séance de communications offertes sur la « combinaison d'enquêtes qu'il a organisée à l'assemblée annuelle de 1999 de l'Institut international de statistique, Leslie a expliqué aux conférenciers que leurs travaux portaient tous sur des aspects différents d'un même problème, qu'ils en soient conscients ou non. La portée de la tâche s'étend aux diverses formes de cumul de données recueillies au moyen d'échantillons successifs, ainsi qu'aux moyens de combiner les données provenant de divers pays afin de produire des statistiques pour de plus grandes entités, comme l'Union européenne. Kish (2001) soutient que ces problèmes sont fondamentalement les mêmes que celui de la combinaison d'information provenant d'expériences différentes (Cochran 1937, 1954).

3. LE QUESTIONNAIRE DÉTAILLÉ DU RECENSEMENT ET LES OPTIONS INTERCENSITAIRES

Le « questionnaire détaillé » du recensement décennal est la source principale de données internationales sur les

caractéristiques de la population et du logement aux États-Unis. Les estimations du nombre de personnes et d'unités de logement sont calculées d'après le « questionnaire abrégé » du recensement auquel répondent tous les ménages. Le questionnaire détaillé, pour lequel le taux global d'échantillonnage est de un sur six, fournit des estimations précises, détaillées (Le terme « précises » a trait aux erreurs d'échantillonnage et le terme « détaillées » signifie que les estimations sont données pour un grand nombre de domaines démographiques dans un domaine géographique), de diverses caractéristiques démographiques et économiques pour les divers États, les grandes villes, ainsi que les grands comtés ou groupes de comtés. Il permet de produire des estimations utiles, quoique moins précises et moins détaillées, pour des régions encore plus petites, comme les petites villes et les réserves indiennes, ainsi que pour les secteurs de recensement, qui comptent, en moyenne, 4 000 habitants. Pour les unités gouvernementales les plus petites, on utilise des taux d'échantillonnage plus élevés, pouvant atteindre un sur deux pour les zones les plus petites, de sorte que l'on puisse produire des estimations utilisables pour ces domaines. Afin de compenser le taux d'échantillonnage plus élevé utilisé pour ces régions, celui appliqué aux secteurs de recensement les plus grands est de un sur huit.

Entre les recensements, les programmes statistiques du gouvernement fédéral fournissent assez peu de renseignements sur les caractéristiques de la population à un niveau international. Les dénombremens de base du recensement

sont mis à jour grâce à un programme intercensitaire d'estimations démographiques, mais les données sur d'autres caractéristiques démographiques et économiques proviennent principalement d'enquêtes nationales. Dans le cas de la *Current Population Survey* (CPS), qui est l'enquête américaine mensuelle sur la population active, le taux d'échantillonnage est d'environ 1 sur 1 000 et le chevauchement entre les unités d'échantillonnage est considérable d'un mois à l'autre, de sorte que l'échantillon ne peut être cumulé profitablement au cours du temps comme cela est possible dans le cas d'échantillons successifs avec renouvellement complet. Un supplément à la CPS est réalisé une fois par an en mars afin de recueillir des renseignements supplémentaires pour produire des estimations du revenu et de la pauvreté au niveau de l'État, mais la précision et le niveau de détail démographique de ces estimations sont limités. Certains programmes s'appuient sur des méthodes de modélisation fondées sur des dossiers administratifs pour produire des estimations du chômage, ainsi que du revenu et de la pauvreté pour les petites régions, mais ne permettent pas de produire des estimations pour diverses autres caractéristiques.

La nécessité de recueillir plus fréquemment des renseignements pour les petits domaines (ou « communautés ») est admise depuis longtemps (Häuser 1942; Eckler 1972, page 212; Bonpane 1986). Leslie a reconnu à son ami, Philip Häuser, l'honneur d'avoir proposé un « recensement annuel par sondage » en 1941. Kish (1981) a proposé un plan de sondage à échantillons successifs comme moyen d'atteindre cet objectif et a présenté plusieurs options, y compris un plan de sondage à échantillons successifs pour la CPS. Au lieu de cette proposition, un recensement de milieu de décennie a été approuvé pour 1985, mais n'a jamais été financé. Une proposition visant à doubler la taille de l'échantillon de la CPS ne l'a pas été non plus (Tupek, Waite et Cahoon 1990).

L'intérêt du Census Bureau pour des données intercensitaires sur les caractéristiques de la population s'est ravivé à la suite d'une proposition de « programme décennal de recensement » présentée par Herriot, Bateman et McCarthy (1989). Ce programme aurait permis de recueillir des données pour différents États pour des années différentes; en dernière analyse, la proposition n'a pas été acceptée. Cependant, le plaidoyer dynamique et éloquent de Roger Herriot quant à l'importance et à la valeur éventuelles des données intercensitaires de niveau international a sensibilisé les organismes statistiques fédéraux à la possibilité d'adopter un « nouveau paradigme » pour le cycle décennal de collecte des données. L'échantillonnage a réussi avec renouvellement déterminé pendant cette période où le Bureau a envisagé de nouvelles approches pour le Recensement de 2000 (voir Bonpane 1986).

Le Congrès ayant manifesté un regain d'intérêt pour des données intercensitaires sur les caractéristiques de la population (Meinick 1991; Sawyer 1993), l'option d'une

Les échantillons successifs de Leslie Kish et l'American Community Survey

CHARLES H. ALEXANDER¹

RÉSUMÉ

Leslie Kish a longtemps prôné l'adoption de plans de sondage à « échantillons successifs », ou échantillons à renouvellement complet, avec panels mensuels non chevauchants pouvant être cumulés sur des périodes de diverses longueurs pour des domaines de taille variable. Dans ces conditions, une même enquête peut être utilisée pour atteindre plusieurs objectifs. La nouvelle enquête américaine sur les communautés (*American Community Survey* (ACS)) du Census Bureau est fondée sur un plan de sondage à échantillons successifs de ce genre. Les données qu'elle fournit servent au calcul de moyennes annuelles pour mesurer les variations au niveau de l'État et à celui de moyennes mobiles sur trois ans ou sur cinq ans pour décrire des domaines de plus en plus petits. L'article décrit l'influence exercée par Kish sur l'élaboration de l'American Community Survey, ainsi que certains problèmes méthodologiques d'ordre pratique qu'il a fallu résoudre pour mettre en œuvre le plan de sondage.

MOTS CLÉS : Échantillons successifs; moyennes pluriannuelles; cumuls asymétriques.

1. INTRODUCTION

Tel que défini plus bas, un « plan de sondage à échantillons successifs » permet de créer une enquête unique suffisamment souple pour poursuivre plusieurs objectifs. Le concept a été élaboré par Leslie Kish dans une série d'articles (y compris Kish 1979a, 1979b, 1981, 1983, 1986, 1990, 1997, 1998 et Kish et Verma 1983, 1986) dans lesquels il a énoncé les principes du cumul spatial et temporel d'information. Kish a prôné l'adoption de ce genre de plan de sondage pour atteindre des objectifs divers (Kish 1998), particulièrement dans les pays en voie de développement (Kish 1979b), mais aussi dans le contexte du Recensement des États-Unis (Kish, Lovejoy et Rakow 1981). Son utilisation personnelle des échantillons successifs, qu'il qualifiait alors d'« échantillon continu », remonte au moins à 1958 (Kish 1961; un projet antérieur (Mooney 1956) est mentionné dans Kish (1998)).

L'American Community Survey (ACS), qui est destinée à remplacer l'enquête à « questionnaire détaillé » réalisée habituellement dans le cadre du recensement, s'appuiera sur une forme de plan de sondage à échantillons successifs. Le présent article décrit comment le concept des échantillons successifs est appliqué dans le cas de l'ACS, en tenant compte des objectifs et des considérations opérationnelles propres à l'enquête. Les décisions concernant le plan de sondage de l'ACS illustrent certaines questions que peut soulever l'utilisation d'échantillons successifs en général. Elles illustrent aussi les divers niveaux — conceptuel, personnel et pratique — sur lesquels s'est exercée l'influence de Leslie Kish dans le domaine du développement des enquêtes.

Un plan de sondage à « échantillons successifs » consiste à sélectionner conjointement k échantillons probabilistes (panels) non chevauchants qui représentent chacun $1/k$ de la population complète. À chaque période, les membres d'un panel particulier sont interviewés. Après k périodes, tous les membres de l'échantillon auront été interviewés. Selon la précision exigée, un seul panel de $1/k$ peut suffire à la production de bonnes estimations pour la population dans son ensemble et, éventuellement, pour certains grands domaines. Pour les domaines plus petits, ou pour obtenir une plus grande précision pour les grands domaines, on peut cumuler un nombre variable de panels consécutifs, jusqu'à concurrence de k/F de la population. Un plan de sondage à échantillons successifs pour lequel $k = F$ porte le nom de « recensement continu ». Pour un échantillonnage successif mensuel, il est naturel que F soit égal à un multiple de 12 et que les cumuls soient trimestriels, semi-annuels, annuels ou pluriannuels.

Les « domaines » s'entendent des zones géographiques ainsi que des sous-groupes démographiques. Kish (1987, section 2.3) présente un cadre de référence pour le compromis entre les niveaux de détail géographique et démographique, pour un niveau donné de précision. Un élément encore plus fondamental du concept des échantillons successifs est celui du « cumuli asymétrique » des données, sur des périodes de longueur variable pour des tailles de domaine diverses (Kish 1990, 1998). Ce dernier fut élargi plus tard pour donner une vue des similarités fondamentales du calcul de la moyenne spatiale et de la moyenne temporelle (Kish 1998), ainsi que du calcul de la moyenne sur divers domaines démographiques. La souplesse du plan de sondage à échantillons successifs tient

¹ Charles H. Alexander, U.S. Bureau of the Census, Suitland, Maryland, (U.S.A.) 20233.

- FISHER, R.A. (1926). *Statistical Methods for Research Workers*. London: Oliver & Boyd.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953a). *Sample Survey Methods and Theory, I – Methods and Applications*, New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953b). *Methods and Applications, II – Theory*. New York: John Wiley & Sons, Inc.
- JENSEN, A. (1926). The representative method in practice, *Bulletin of the International Statistical Institute*, 22, pt. 1, 359-439.
- KENDALL, M.G. (1968). Chance, *Dictionary of the History of Ideas*, (Ed. P.P. Wiener), New York: Chas Scribners.
- KIAER, A.W. (1895). The Representative Method of Statistical Surveys, English translation, 1976, Oslo: Statistisk Sentralbyro.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1981). *Using Cumulated Rolling Samples*. Washington DC: Library of Congress.
- KISH, L. (1985). Chance, statistics, sampling, *Journal of Official Statistics*, 1, 35-47.
- KISH, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.
- KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet, *Techniques d'enquête*, 16, 67-86.
- KISH, L. (1994). Multipopulation survey designs, *International Statistical Review*, 62, 167-186.
- KISH, L. (1998). Spacetime variations and rolling samples, *Journal of Official Statistics*, 14, 31-46.
- KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques, *Techniques d'enquête*, 25, 147-158.
- KISH, L., LOVEJOY, W. et RACKOW, P. (1961). A multistage probability sample for continuous traffic surveys, *Proceedings of the American Statistical Association, Section on Social Statistics*, 227-230.
- KRUSKAL, W.H. et MOSTELLER, F. (1979a). Representative sampling, I: Non-scientific literature, *International Statistical Review*, 47, 13-24.
- KRUSKAL, W.H. et MOSTELLER, F. (1979b). Representative sampling, II: Non-scientific literature, *International Statistical Review*, 47, 111-127.
- KRUSKAL, W.H. et MOSTELLER, F. (1979c). Representative sampling, III: The current statistical literature, *International Statistical Review*, 47, 245-265.
- KRUSKAL, W.H. et MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics, *International Statistical Review*, 48, 169-195.
- MOONEY, H.W. (1956). Methodology of Two California Health Surveys, US Public Health Monograph 70, Washington DC: US Government Printing Office.
- NEYMAN, J. (1934). On the different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, 558-625.
- O'MURCHÉARTAIGH, C., et Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: a review, *Bulletin of the International Statistical Institute*, Buenos, 43^{ème} Session, 1, 465-493.
- PORTER, T.M. (1987). The Rise of Statistical Thinking: 1820-1900, Princeton, NJ: Princeton University Press.
- STIGLER, S.M. (1986). History of Statistics, Cambridge: Harvard University Press.
- UNITED NATIONS STATISTICAL OFFICE (1950). The Preparation of Sample Survey Reports, New York: UN Series C No 1; aussi révision 2 en 1964.
- VON MISES, R. (1939). *Probability, Statistics, and Truth*, London: Wm. Hodge and Co.

inférentielles – de deuxième ordre ou plus élevé – sont plus difficiles et diversifiées et sont expliquées à part dans la documentation. On remarquera que l'échantillonnage probabiliste est la situation spéciale (et souvent souhaitée) où tous les k_i sont égaux à 1.

L'autre objection à la terminologie d'échantillonnage probabiliste est plus théorique et philosophique puisqu'elle porte sur le mot « connu » dans sa définition. Ce mot semble sous-entendre une croyance. Des auteurs, à partir des classifications comme John Venn et M. G. Kendall jusqu'à des bayésiennes modernes comme Dennis Lindley – et d'autres aux deux extrémités – ont clairement attribué la « probabilité » à des états de croyance et le « hasard » à des fréquences produites par des phénomènes objectifs et des opérations mécaniques. Par conséquent, notre insistance sur des opérations, comme des générateurs de nombre aléatoires, devrait sous-entendre l'expression « échantillonnage au hasard ». Toutefois, je n'en ai pas observé l'emploi et, de plus, il pourrait déboucher sur un problème philosophique : l'utilisation appropriée de bonnes tables de nombres aléatoires sous-entend une croyance dans leurs probabilités « espérées ». Je n'ai passé que très peu de temps à ces problèmes et à des discussions agréables avec seulement un très petit nombre de collègues, qui ont cependant été d'accord avec moi. Je serais heureux de participer à d'autres discussions ou de recevoir des suggestions et des corrections.

6. SUR DES SUJETS APPARENTÉS

Nous avons demandé la reconnaissance de nouveaux paradigmes pour six aspects de l'échantillonnage d'enquête, nous soulignons ici le contraste entre l'échantillonnage et d'autres méthodes apparentées. Les méthodes d'enquête donnent le choix et la définition des variables, des méthodes des exploitants agricoles possédant plusieurs terrains. Des exemples de $k_i < 1$ surviennent lors du prélèvement d'un seul adulte au sein de ménages, ou de logements parmi des immeubles. Dans ces exemples, il arrive souvent que k_i puisse facilement être vérifié, et il est alors plus facile, plus pratique et plus économique de recourir à des coefficients de pondération que d'essayer d'obtenir la valeur $1/F$ pour tous les éléments. Ces problèmes sont décrits dans des ouvrages et des articles.

Dans la plupart des cas, il est plus pratique et moins coûteux d'accepter les probabilités variables et de les contrebalancer par la pondération des valeurs espérées $1/k_i$ ou k_i que d'ajouter une autre étape du prélèvement. Par conséquent, pour paraphraser l'échantillonnage probabiliste : l'échantillonnage espéré assure pour chaque élément de la population ($i = 1, 2, \dots, N$) un nombre espéré positif connu de prélèvements ($k_i/F > 0$). Dans la pratique, on utilise ces procédures pour des statistiques descriptives (du premier ordre) où les variables k_i ou $1/k_i$ ne sont ni élevées ni fréquentes. Les traitements des statistiques

La plus importante des deux concerne les situations pratiques qui se présentent fréquemment lorsque nous devons faire un choix entre l'échantillonnage probabiliste et l'échantillonnage espéré. Ces situations se produisent souvent lorsque le taux de prélèvement pratique et facile des unités de listage de $1/F$ donne non seulement la probabilité unique $1/F$ pour les éléments, mais aussi une probabilité avec la fraction variable k_i/F pour le $i^{\text{ème}}$ élément ($i = 1, 2, \dots, N$) et avec $k_i > 0$. Des exemples de $k_i > 1$, habituellement un petit nombre entier, surviennent avec des listes avec répétitions, des bases de prélevement doubles, les résidences secondaires de ménages, des populations mobiles et nomades, ou des exploitants agricoles possédant plusieurs terrains. Des exemples de $k_i < 1$ surviennent lors du prélèvement d'un seul adulte au sein de ménages, ou de logements parmi des immeubles. Dans ces exemples, il arrive souvent que k_i puisse facilement être vérifié, et il est alors plus facile, plus pratique et plus économique de recourir à des coefficients de pondération que d'essayer d'obtenir la valeur $1/F$ pour tous les éléments. Ces problèmes sont décrits dans des ouvrages et des articles.

5. ÉCHANTILLONNAGE ESPÉRÉ

L'échantillonnage probabiliste assure, pour chaque élément de la population ($i = 1, 2, \dots, N$) une probabilité de prélèvement positive connue ($P_i > 0$). Cette assurance doit se fonder sur une procédure mécanique de prélèvement au hasard plutôt que seulement sur des hypothèses, des croyances ou des modèles au sujet des distributions probabilistes. La procédure de randomisation exige une opération physique pratique qui correspond de très près (ou exactement) au modèle probabiliste (Kish 1965). La plupart des ouvrages sur l'échantillonnage d'enquête présentent une affirmation de ce genre, et j'y crois encore. Toutefois, cette définition et ses exigences souffrent de deux objections

4. REGROUPEMENT D'ÉCHANTILLONS DE POPULATION

Les échantillons de populations nationales représentent toujours des sous-populations (domaines) qui affichent des caractéristiques d'enquête parfois légèrement différentes, parfois fortement différentes. On peut distinguer ces sous-classes dans l'échantillon avec plus ou moins de peine. En premier lieu, on peut facilement séparer les échantillons de provinces lorsque leur prélevement fait l'objet d'une opération distincte. En deuxième lieu, on peut aussi distinguer, et parfois les utiliser comme des estimations poststratifiées, des sous-classes par âge, sexe, profession et niveau de scolarité. En troisième lieu, cependant, les sous-classes de finies par leurs caractéristiques sociales, psychologiques et comportementales peuvent parfois se prêter difficilement à une différenciation même si elles présentent le plus de lien avec les variables d'enquête. Par conséquent, nous admettons que les échantillons nationaux ne sont pas de simples agrégations de personnes provenant d'une population à DII, mais plutôt des combinaisons de sous-classes provenant de sous-populations affichant des caractéristiques diversifiées. La composition de populations nationales elle sert aussi d'exemple aux deux types de combinaison qui suivent. De plus, ces remarques s'adressent à des combinaisons non seulement d'échantillons nationaux, mais aussi de villes, d'institutions, d'établissements, etc.

On observe depuis quelques années deux types de plan de sondage qui exigent des efforts au-delà de ceux que commandent de simples échantillons nationaux : a) les échantillons périodiques, et b) les plans de populations multiples. Ces deux plans sont uniquement de facture récente, car ils devaient faire appel à trois types de ressources financières et politiques : 2. des ressources techniques institutionnelles suffisantes dans les bureaux statistiques nationaux ; 3. de nouvelles méthodes. Pour ces deux types de plan, nous devons distinguer les besoins des méthodes d'enquête (définitions, variabilités, mesures), qu'il faut harmoniser et normaliser, et ceux liés aux plans de sondage, qui peuvent être librement conçus pour s'adapter à des situations nationales (et même provinciales), à condition de reposer sur un plan probabiliste (Kish 1994). Les deux types de plan ont été conçus d'abord et avant tout à des fins de comparaisons, c'est-à-dire des comparaisons périodiques et des comparaisons multinationales respectivement. Toutefois, de nouvelles utilisations sont aussi apparues : des « échantillons à renouvellement complet » et des cumuls multinationaux, respectivement. Chaque type de cumulo a fait face à une très forte opposition et il doit donner lieu à une nouvelle vision, à un nouveau paradigme.

On a utilisé à quelques reprises les « échantillons à renouvellement complet » dans des situations locales (Mooney 1956; Kish, Lovejoy et Rackow 1961). On les a ensuite proposés plusieurs fois à la place d'échantillons

annuels nationaux et aussi comme remplacement possible de recensements décennaux (Kish 1981, 1990). Mais c'est d'abord et avant tout le Census Bureau des États-Unis (Alexander 1999; Kish 1990) qui les utilise maintenant à la place d'échantillons pour des recensements nationaux. En recommandant cette nouvelle méthode, je dois habituellement répondre à des questions au sujet des moyennes des échantillons périodiques : « Comment pouvez-vous établir la moyenne d'échantillons lorsque ces échantillons varient entre les périodes ? » À mon avis, plus la variabilité est prononcée, moins grande devrait être la dépendance à une période unique, que la variation soit monotone, cyclique ou aléatoire. Je constate donc deux visions ou paradigmes contrastants. Très souvent, l'opposition disparaît après deux jours de discussions et de réflexion.

« Par exemple, le revenu annuel est une agrégation facilement acceptée, non seulement pour les revenus stables, mais aussi pour des professions présentant des variations marquées (saisonnnières ou irrégulières). Il se révélera plus facile d'accepter des moyennes d'échantillons hebdomadaires pour des statistiques annuelles que des moyennes d'échantillons cumulés sur dix ans. Néanmoins, beaucoup d'investisseurs dans les fonds communs d'actions préfèrent utiliser la moyenne mobile des gains sur dix ans ou cinq ans (malgré son caractère périmé) plutôt que les gains actualisés de l'année précédente (avec leurs variations « aléatoires » risquées. Dans la planification d'un pique-nique, la température « normale » moyenne sur cinquante ans au chiffre exact de température de l'année précédente. Il y a de nombreux exemples semblables de méthodes perfectionnées d'établissement de la moyenne sur de longues périodes auquel recourt le public « naïf ». Ce public, et les décideurs, s'accommoderaient aussi rapidement des échantillons à renouvellement complet si on leur en donnait la chance » (Kish 1998)

Tout comme pour les échantillons à renouvellement complet, les échantillons combinés de populations multiples ont aussi rencontré une opposition : les frontières nationales représentent différentes étapes de leur évolution, ainsi que des différences au niveau des lois, des langues, des cultures, des coutumes, des religions et des comportements. Comment peut-on alors les combiner ? Cependant, nous trouvons souvent des utilisations et des sens à des moyennes continentales ; par exemple, les taux de naissance et de décès en Europe, ou les taux en Amérique du Sud, en Afrique sub-saharienne ou en Afrique occidentale. C'est parfois aussi le cas de la naissance, de décès et de croissance à l'échelle mondiale. Cependant, n'ayant pas fait

« superpopulations ». La simplicité à cause des complexités des ces modèles est nécessaire à la base de chacun de ces modèles. Toutes les sciences ont besoin de modèles simples à leur début : par exemple, une orbite parfaitement circulaire pour les planètes, ou $d = gt^2/2$ pour les objets en chute libre dans un milieu dénué de friction. Toutefois, ces modèles ne répondent pas à la complexité du monde physique réel. De même, aucune *population*, fût-elle humaine, animale, végétale, physique, chimique ou biologique, ne comporte des éléments indépendants. Les modèles indépendants simples peuvent servir assez bien aux petits *échantillons*; et on a souvent utilisé comme exemple (précieux parce que rare) la distribution de Poisson des décès par ruades dans l'armée prussienne au cours d'une période de 43 ans (Fisher 1926).

Il y a aussi eu des tentatives de construire des populations théoriques d'éléments à DII: la plus connue est sans doute le « collectif » classique de Von Mises (1931); cependant, elles ne correspondent pas à des populations réelles. Toutefois, on a confectionné, avec beaucoup d'efforts, des tables de nombres aléatoires qui ont subi avec succès tous les tests. On les a largement utilisées dans des expériences modernes et dans des sondages. La *réplication* et la *randomisation* sont deux des notions les plus fondamentales de la statistique moderne qui s'inspirent du concept des populations.

Le simple concept d'une population composée d'éléments indépendants ne décrit pas assez bien les distributions complexes (dans l'espace, dans le temps et dans les classes) des éléments. Le groupement et la stratification sont des noms communs de complexités omniprésentes. De plus, il semble impossible de construire des modèles qui pourraient mieux décrire des populations réelles. Les distributions sont beaucoup trop complexes, outre qu'elles soient aussi différentes pour chacune des variables de l'enquête. On a examiné ces complexités et différences qu'on présente maintenant sous forme de milliers de calculs « d'effets du plan ».

Paradigme pour faire face aux complexités des populations de tous genres comportant un grand nombre de variables d'enquête et une liste croissante de statistiques d'enquête. On a donc conçu des plans de prélèvement et des multide variance robustes susceptibles de convenir à une multitude de plans de sondages, ce qui a donné naissance à la nouvelle discipline de l'échantillonnage d'enquête. Le calcul des « effets du plan » a fait ressortir l'existence, l'ampleur et la variabilité des effets attribuables aux complexités des distributions, non seulement pour les moyennes mais aussi pour les relations à variables multiples, comme les coefficients de régression. La longue période de désaccord entre les échantillonneurs d'enquête et les économétriciens témoigne du besoin de se doter d'un nouveau paradigme.

3. POPULATIONS COMPLEXES

La force, l'ampleur et la durée de la résistance manifestée à l'égard des notions de l'utilisation de la population probabiliste appliquée à des populations de base signifient qu'il s'agit d'un nouveau paradigme qui commande de la part du public et des spécialistes une nouvelle vision.

La théorie et sur de vastes applications. entre 1949 et 1954 amorcèrent un flot d'articles sur la population probabiliste. Cinq ouvrages d'influence parus (Unies 1950) renverser le courant en faveur de l'échantillonnage probabiliste appliqué à des populations de base signifient qu'il s'agit d'un nouveau paradigme qui commande de la part du public et des spécialistes une nouvelle vision.

La force, l'ampleur et la durée de la résistance manifestée à l'égard des notions de l'utilisation de la population probabiliste appliquée à des populations de base signifient qu'il s'agit d'un nouveau paradigme qui commande de la part du public et des spécialistes une nouvelle vision.

Il s'écoulera un demi-siècle avant que le document de Kiser ne débouche sur une vaste acception de l'échantillonnage d'enquête. Outre la négligence et la résistance passive, on pouvait aussi observer une forte opposition de la part des bureaux statistiques nationaux, qui se méfiaient des méthodes d'échantillonnage destinées à remplacer les chiffres complets des recensements. D'aucuns préféraient même la « méthode de la monographie », qui présentait des chiffres complets provenant d'une province ou d'un district national prélevé de façon aléatoire (O'Muircheartaigh et Wong 1981). Outre le milieu politique, de nombreux opposants provenaient du milieu académique, dont des statisticiens. La publication du rapport de la Commission statistique des Nations Unies, sous la direction de Mahalanobis et Yates (Bureau statistique des Nations Unies 1950) renverser le courant en faveur de l'échantillonnage probabiliste. Cinq ouvrages d'influence parus entre 1949 et 1954 amorcèrent un flot d'articles sur la théorie et sur de vastes applications.

L'échantillonnage d'enquête se distingue notamment par un prélevement probabiliste strict d'une base de population, qui permet d'obtenir des inférences de l'échantillon pour les applications complexes des éléments dans les populations représentées des problèmes encore plus importants et difficiles. Ces complexités contrastent vivement avec le modèle d'indépendance simple que présupposent, explicitement ou implicitement, presque toute la théorie statistique et toutes les statistiques mathématiques.

L'hypothèse d'observations indépendantes, ou non corrélées, de variables ou d'éléments est sous-jacente aux statistiques mathématiques et à la théorie de la distribution. Nous n'avons pas besoin ici de faire une distinction entre des variables aléatoires à distribution indépendante et identique (DII) et une « échangeabilité », et des opérations mécaniques du prélevement aléatoire. C'est sur quoi seront basées les inférences statistiques entre les statistiques de la population (paramètres) (Hansen, Hurwitz et Madow 1953a, 1953b). Cette insistance sur des inférences reposant sur des prélevements de populations de base représente un paradigme différent des approches imprécises ou fondées sur un modèle qu'utilisent la plupart des analyses statistiques.

Nouveaux paradigmes (modèles) pour l'échantillonnage probabiliste

LESLIE KISH¹

1. UN NOUVEAU PARADIGME POUR LA STATISTIQUE

À plusieurs endroits, j'aborde de nouvelles notions liées à divers aspects de l'échantillonnage, mais je ne sais trop si je devrais les appeler de nouveaux paradigmes ou de nouveaux modèles ou simplement de nouvelles méthodes. À cause de mon incertitude et de mon manque de confiance, je prie le lecteur de choisir l'expression qui lui sied le mieux. Je préfère que ce terme ne devienne pas un obstacle à notre compréhension mutuelle.

L'échantillonnage est à la fois une branche et un outil de

la statistique. En 1810, après la naissance de certaines

sciences comme l'astronomie, la chimie et la physique,

Quelet (Porter 1987; Stigler 1986) instituait le domaine de

la statistique sous forme de paradigme. « À la fin du

XVII^e siècle, les études philosophiques sur la cause et le

hasard... ont commencé à se rapprocher... aux XVIII^e et

XIX^e siècles, on en vient de plus en plus à admettre que,

contrairement à des groupes de personnes, des groupes

d'événements sont susceptibles d'obéir à des lois »

(Kendall 1968). Les régularités prévisibles, significatives et

utiles constatées dans le comportement de groupes de

populations composés d'individus non prévisibles furent

nommées « statistiques » et constituèrent dès lors une grande

découverte.

Quelet et d'autres calculèrent donc des taux de natalité,

des taux de décès, des taux de suicide, des taux d'homicide,

des taux d'assurance, etc., à l'échelle nationale (et autres)

Ces statistiques sont fondamentales à des domaines comme

la démographie et la sociologie. Plus tard au XIX^e siècle,

Frances Galton et Karl Pearson, puis Maxwell, intro-

duisirent les notions de statistiques en biologie et en

physique respectivement, puis la théorie et les applications

tirent un grand pas.

La statistique et les statisticiens s'intéressent aux effets

du hasard sur des données empiriques. Les mathématiques

du hasard ont été mises au point plusieurs siècles

auparavant pour les jeux de hasard ainsi que pour les

calculs des erreurs d'observation en astronomie. Des

données avaient aussi été compilées pour le commerce, les

banques et les administrations publiques. Toutefois, un

nouveau paradigme, une nouvelle vision théorique s'impose

si l'on désire combiner le hasard à des données réelles. Pour

cette raison, étant le produit de la maturité du

développement humain, la science statistique et ses diverses

branches arrivèrent sur le tard dans l'histoire et dans les

universités (Kish 1985).

Le concept le plus fondamental de la statistique repose

sur des populations d'individus choisis au hasard. C'est le

fondement des théories de la distribution, des inférences, de

la théorie de l'échantillonnage, du plan expérimental, etc.

De plus, le paradigme de la statistique s'écarte fondamen-

talement de la vision déterministe de la cause et de l'effet

ainsi que des relations précises observées dans les autres

sciences et les mathématiques.

2. LE PARADIGME DE L'ÉCHANTILLONNAGE

Publiée presque cent ans après la naissance de la

statistique, il y a maintenant plus d'un siècle, une impor-

tante monographie portant le titre *The Representative*

Method est généralement acceptée comme le point de

départ de l'échantillonnage moderne (Kiaer 1895).

L'expression a été utilisée depuis lors dans plusieurs

documents ayant fait autorité (densen 1926; Neyman 1934;

Kruskal et Mosteller 1979a, 1979b, 1979c, 1980). Ces

derniers auteurs admettent que le terme « représentatif » a

été accolé à un si grand nombre de méthodes spécifiques et

avec des sens tellement différents qu'il ne désigne plus

aujourd'hui une méthode particulière. Toutefois, dans la

forme où Kiaer l'a employé, et comme on l'utilise encore de

façon générale, le terme désigne le prélèvement d'un

échantillon destiné à représenter une population spécifique

dans l'espace, dans le temps et sous d'autres définitions,

afin de pouvoir obtenir de l'échantillon des inférences

statistiques se rapportant à cette population. Pour cette

raison, un échantillon représentatif national doit reposer sur

éléments de la population nationale, et non pas seulement

au sein d'un quelconque domaine arbitraire comme une

ville ou une province « typique », ou d'un sous-ensemble,

quoiqu'il soit défini ou non.

Méthode scientifiquement acceptée pour l'échan-

tilonnage d'enquête, l'échantillonnage probabiliste assure

des probabilités de prélèvement positives connues pour

chacun des éléments de la population de base. La base

fournit l'équivalent de listes d'unités d'échantillonnage à

chaque étape du prélèvement. La base d'échantillonnage de

la population complète est nécessaire pour effectuer les

L'impression de cet article a été autorisée avec gentillesse par Rhea Kish, 1050 Wall St. #9A, Ann Arbor, MI 48105, courrier électronique: rheak@umich.edu.

- KISH, L. (1981). Using Cumulated Rolling Samples. Washington: Library of Congress.
- KISH, L. (1982). Design effects. *Encyclopedia of Statistics*, New York: John Wiley & Sons, Inc.
- KISH, L. (1983). Data collection for details over space and time. *Statistical Methods and the Improvement of Data Quality*, (Ed. T. Wright), New York: Academic Press.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.
- KISH, L. (1987a). *Statistical Design for Research*. New York: John Wiley & Sons.
- KISH, L. (1987b). Discussion. *Small Area Statistics*, (Ed. R. Platek). New York: John Wiley & Sons, Inc.
- KISH, L. (1988). Plans de sondage à usages multiples. *Techniques d'enquête*, 14, 19-33.
- KISH, L. (1989). *Sampling Methods for Agricultural Surveys*. Rome: FAO.
- KISH, L. (1990). Recenseusement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 66-77 et 99-100.
- KISH, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339-347.
- KISH, L. (1995a). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.
- KISH, L. (1995b). *Questions/answers from the survey statistician 1978-1994*. Librairie: International Association of Survey Statisticians.
- KISH, L. (1996). Developing samplers for developing countries. *International Statistical Review*, 64, 143-162.
- KISH, L. (1997). Periodic and rolling samples and censuses. *Statistics and Public Policy*, (Ed. B.D. Spencer). New York: Oxford University Press.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.
- KISH, L. (1999). Le cumul ou la combinaison d'enquêtes démographiques. *Techniques d'enquête*, 25, 147-158.
- KISH, L. (2002). Combining multi-population surveys. *Journal of Statistical Planning and Inference*, 102, 109-118.
- KISH, L., et ANDERSON, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 24-34.
- KISH, L., et FRANKEL, M. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, B*, 36, 1-37.
- KISH, L., FRANKEL, M.R. et VAN ECK, M. (1972). *SEPP: Sampling Error Programs Package*. Ann Arbor: Institute for Social Research.
- KISH, L., FRANKEL, M.R., VERMA, V. et KACIROTI, N. (1995). Effets du plan de sondage sur les $(P_i - P_j)$ corrigés. *Techniques d'enquête*, 21, 131-139.
- KISH, L., GROVES, R.M. et KROTKI, K. (1976). Sampling Errors For Fertility Surveys. Occasional Paper No. 17, World Fertility Survey.
- KISH, L., et HESS, I. (1958). On noncoverage of sample dwellings. *Journal of the American Statistical Association*, 53, 509-524.
- KISH, L., et HESS, I. (1959a). On variances of ratios and their differences in multi-stage samples. *Journal of the American Statistical Association*, 54, 416-446.
- KISH, L., et HESS, I. (1959b). A replacement procedure for reducing the bias of nonresponse. *The American Statistician*, 13, 17-19.
- KISH, L., et LANSING, J.B. (1954). Response error in estimating the value of homes. *Journal of the American Statistical Association*, 49, 520-538.
- KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- KISH, L., et VERMA, V. (1986). Complete censuses and samples. *Journal of Official Statistics*, 2, 381-96.
- PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (small domains). *International Statistical Review*, 48, 3-18.

4. CONCLUSION

Leslie Kish est un géant dans le domaine de l'échantillonnage. Ses travaux impressionnants lui ont valu de nombreux honneurs. Ceux-ci incluent, entre autres, la fonction de président de l'Association internationale des statisticiens d'enquêtes de 1983 à 1985, la fonction de président de l'American Statistical Association en 1978 (voir Kish, 1978 pour son discours du président ayant pour thème « Chance, Statistics and Statisticians », les titres d'Honorary Fellow de l'Institut international de statistique, d'Honorary Member de la Hungarian Academy of Sciences, de Fellow de l'American Association for the Advancement of Science, de Fellow de l'American Academy of Arts and Sciences, de lauréat de l'American Statistical Association's Samuel L. Wilks Award en 1997, de lauréat du Mindel Shep Award de la Population Association of America en 1998, de lauréat du Methodology Award de l'American Sociological Association en 1989, ainsi que l'octroi de grades honoris causa de l'Université de Bologne, de l'Athens University of Economics and Business et de l'Eotvos Lorand University de Budapest.

Pourtant, Kish a toujours eu les pieds sur terre et était ouvert à tous. Il discutait avec enthousiasme de nombreux sujets, y compris de sport, d'art, de littérature, de politique, de philosophie et de science. Il se souciait sans cesse d'améliorer les conditions de la population mondiale. Il s'intéressait tout spécialement aux jeunes et l'une de ses phrases préférées était « Restez jeune en étant curieux et aimez de jeunes amis ». Sans aucun doute, sa personnalité attachante a joué un rôle important dans le succès avec lequel il a fait la promotion de l'utilisation de bonnes méthodes d'échantillonnage partout dans le monde.

ALEXANDER, C. H. (2002). Les échantillons successifs de l'enquête, 28, 37-44.

ANDERSON, D.W., KISH, L. et CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

ANDERSON, D.W., KISH, L. et CORNELL, R.G. (1980). On stratification, grouping, and matching. *Scandinavian Journal of Statistics*, 7, 61-66.

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

FELLEGI, I.P. (2000). Leslie Kish – Une vie de dévouement. *Techniques d'enquête*, 26, 133-134.

FRANKEL, M., et KING, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 65-87.

BIBLIOGRAPHIE

- GOODMAN, R., et KISH, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- KALSBEEK, W.D. (1973). A Method for Obtaining Local Postcensal Estimates for Several Types of Variables. Thèse de doctorat, University of Michigan.
- KALTON, G., et KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, 13(16), 1919-1939.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KISH, L. (1952). On the Differentiation of Ecological Units. Thèse de doctorat, University of Michigan.
- KISH, L. (1954). Differentiation in metropolitan areas. *American Sociological Review*, 19, 388-398.
- KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review*, 22, 1954-1965.
- KISH, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24, 328-338.
- KISH, L. (1961a). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*, 4, 201-209.
- KISH, L. (1961b). Efficient allocation of a multi-purpose sample. *Econometrica*, 29, 363-385.
- KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- KISH, L. (1965a). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1965b). Selection techniques for rare traits. *Genetics, Epidemiology, and Chronic Diseases*, Public Health Service Publication, No. 1173.
- KISH, L. (1965c). Sampling organizations and groups of unequal sizes. *American Sociological Review*, 20, 564-572.
- KISH, L. (1968). Standard errors for indexes from complex samples. *Journal of the American Statistical Association*, 63, 512-529.
- KISH, L. (1969). Design and estimation for subclasses, comparisons, and analytical statistics. *New Developments in Survey Sampling*, (Eds. N.L. Johnson et H. Smith). New York: John Wiley & Sons, Inc.
- KISH, L. (1975). Representation, randomization and control. *Quantitative Sociology*, (Eds. H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon et V. Capecchi). New York: Academic Press.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, A*, 139, 80-95.
- KISH, L. (1978). Chance, statistics, and statisticians. *Journal of the American Statistical Association*, 73, 1-6.
- KISH, L. (1979a). Samples and censuses. *International Statistical Review*, 47, 99-109.
- KISH, L. (1979b). Rotating samples instead of censuses. *Asian and Pacific Census Forum* (East-West Center, Honolulu), 6, 1-13.
- KISH, L. (1980). Design and estimation for domains. *The Statistician*, 29, 209-222.

uniquement sur le biais, comme le faisaient la plupart des auteurs ayant traité du sujet.

3. AUTRES CONTRIBUTIONS

Les travaux fondamentaux et de grande portée de Kish dans le domaine de la méthodologie de la statistique d'enquête sont d'une importance considérable. Toutefois, les efforts qu'il a déployés pour promouvoir l'utilisation de méthodes d'échantillonnage probabiliste valides partout dans le monde le sont peut-être plus encore.

Naturellement, les écrits de Kish, qui mettent l'accent sur l'adoption de bonnes méthodes pratiques, ont contribué à l'usage généralisé des méthodes d'échantillonnage probabiliste. Ces trois ouvrages intitulés *Survey Sampling* (Kish 1965a), *Statistical Design for Research* (Kish 1987a) et *Statistical Methods for Agricultural Surveys* (Kish 1989) sont tous fort précieux à cet égard, tout comme le sont ses articles magistraux à l'intention des spécialistes des sciences sociales.

Kish a toujours eu la vocation d'aider les pays en voie de

développement et les pays en transition, et nombre de ses activités en témoignent. Il a été consultant en échantillonnage pour la *World Fertility Survey* de 1973 à 1983 et a donné des consultations dans de nombreux pays. Il s'occupait de l'organisation d'un programme de formation à l'intention des statisticiens étrangers et il a rédigé certains articles s'adressant spécialement aux statisticiens des pays en voie de développement. Ainsi, *Statistical Methods for Agricultural Surveys* a été écrit pour la *Food and Agricultural Organization*, particulièrement à l'intention de ces pays. De 1978 à 1994, il a contribué à la rubrique *Questions/Réponses* du bulletin intitulé *Statistical Methods for Agricultural Surveys* de l'Association internationale des statisticiens d'enquête. Il y donnait d'excellents conseils au sujet de nombreux problèmes pratiques d'échantillonnage qui se posent fréquemment, mais qui ne sont pas toujours bien traités dans les publications. La rubrique était jugée si utile que l'AISS a publié l'ensemble complet de questions et de réponses dans un volume spécial (Kish 1995b).

Kish était, à juste titre, très fier du programme intensif de deux mois sur l'échantillonnage à l'intention des statisticiens étrangers (*Sampling Program for Foreign Statisticians*) qu'il avait créé au *Survey Research Center* en 1961. Aujourd'hui, le SPFS a formé plus de 500 statisticiens d'enquêtes en provenance de 105 pays. Il est significatif que Kish ait choisi « Développement d'échantillonneurs pour les pays en voie de développement » comme thème de la *Morris Hansen Memorial Lecture* de 1994 (Kish 1996). Pour assurer le maintien de ce programme important, lors d'une cérémonie commémorant les 90 ans de Kish, l'Université du Michigan a créé le *Leslie Kish International Fellows Fund*. De toutes ses réalisations, le SPFS était celle qui lui procurait le plus grand plaisir.

comme les hôpitaux et les patients. Si le nombre de patients varie considérablement selon l'hôpital, la production de bonnes estimations au niveau de l'hôpital et au niveau des patients pose un problème. Kish (1965c) a examiné ce problème et précisé les éléments qui entrent en jeu.

Erreurs non dues à l'échantillonnage. Kish a reconnu non dues à l'échantillonnage sur la qualité des estimations produites par sondage. Au tout début de sa carrière, il a étudié, en collaboration de Jack Lansing, les erreurs de réponse lors de la déclaration de la valeur du domicile faite par les répondants en comparant ces déclarations aux estimations faites par des évaluateurs professionnels (Kish et Lansing 1954). Dans son étude de la variance due à l'intervieweur, en s'inspirant de la théorie de l'échantillonnage en grappes, il a mesuré la variance due à l'intervieweur au moyen du coefficient de corrélation intraclass

et déterminé le nombre optimal d'interviews par intervieweur en se fondant sur un modèle de coût de l'échantillonnage en grappes simple (Kish 1962). Avec Irene Hess, il a réalisé une étude de la non-couverture dans le cas d'échantillons ascendants de logements. L'étude, lancée parce que le taux de non-couverture des enquêtes du SRC était de 10 % à cette époque-là, a permis d'apporter des améliorations qui l'ont réduit à environ 3 % (Kish et Hess 1958). Toujours avec Irene Hess, il a introduit une méthode novatrice de remplacement des personnes non contactées lors d'une enquête qui consistait à les substituer par des personnes non contactées d'une enquête similaire antérieure (Kish et Hess 1959b). En ce qui concerne les scénarios d'imputation stochastiques, Kish a été l'un des premiers adeptes de la répétition des imputations en vue de réduire la variance d'imputation, méthode qu'il a baptisée à l'époque *imputation à répétitions répétées*, ou *RRIIP* pour repealed replication imputation procedure, et qui est connue aujourd'hui sous le nom d'imputation fractionnaire (Katon et Kish 1984).

Études par observation. Au début de sa carrière, Kish (1959) a rédigé un article cité très fréquemment sur la conception d'études destinées à étudier les liens de causalité, particulièrement des études non randomisées. Dans ses écrits à ce sujet, il s'est appuyé sur son expérience en échantillonnage, comme celle concernant le lien entre la stratification et l'appariement (Anderson, Kish et Cornell 1980). Ses travaux ont mené à la publication de l'ouvrage intitulé *Statistical Design for Research* (Kish 1987a) dans lequel il compare l'utilisation d'enquêtes, d'expériences et d'études par observation pour étudier les effets de causalité en fonction des trois R : réalisme, randomisation et représentativité (voir aussi Kish 1975). Il a également souligné l'importance qu'il y a à évaluer le biais et la variance pour mesurer les effets de causalité, au lieu de se concentrer

Martin Frankel, étudiant de doctorat dont il avait la supervision, il a également étudié la gamme des statistiques pour lesquelles il était possible de calculer les erreurs d'échantillonnage dans le cas des plans de sondage complexes (Kish et Frankel 1970, 1974). Ces travaux de recherche d'une très grande importance ont mené au développement, à l'application et à l'évaluation de la méthode des répétitions équilibrées répétées, ou BRR pour balancer repeated replication, et de la méthode des répétitions jackknife répétées, ou JRR pour jackknife repeated replication, pour estimer la variance. Il a aussi proposé une définition des paramètres de population estimés d'après des statistiques analytiques fondées sur des données d'enquête dans le contexte des populations finies.

Enquêtes polyvalentes. Les articles sur l'échantillonnage traitent principalement de l'élaboration d'un plan d'échantillonnage efficace pour estimer un seul paramètre d'une population. Kish a reconnu les limites de cette approche, puisque presque toutes les enquêtes ont un caractère polyvalent. Il a rédigé plusieurs articles importants traitant des enquêtes polyvalentes et a produit des plans de sondage de compromis efficaces qui fournissent des estimations non seulement pour l'ensemble de la population, mais aussi pour divers domaines (Kish 1961b, 1969, 1976; Anderson, Kish et Cornell 1976; Kish et Anderson 1978; Kish 1980; Kish 1988). Ces dernières années, il avait étendu son étude pour s'intéresser aux enquêtes à populations multiples (par exemple, Kish 1999, 2002).

Estimation régionale. Afin de produire des estimations par domaine, Kish (1980, 1987a, 1987b) a classifié ces derniers en grands, petits et mini domaines et en éléments rares. Pour les grands domaines, on peut produire les estimations avec les données d'enquête en servant d'estimateurs types basés sur l'échantillon, particulièrement si ce dernier est conçu pour enquêter un nombre suffisant d'unités appartenant aux domaines en question. Toutefois, la taille de l'échantillon de la plupart des enquêtes empêche de produire des estimations suffisamment précises pour les petits ou les mini domaines qui comprennent moins de dix, un dixième de la population. Pourtant, comme Kish l'a compris rapidement, la demande d'estimations à jour pour de petits domaines, particulièrement les petites régions géographiques, allait augmenter. Sa sensibilisation à ce problème est à l'origine des travaux qu'il a menés dans deux domaines connexes.

Si la taille de l'échantillon d'une enquête est trop petite pour que l'on puisse produire des estimations régionales, on peut recourir à des modèles statistiques pour produire des estimations indirectes. De nombreux travaux portant sur les techniques d'estimation régionale au moyen de modèles ont été réalisés ces dernières années. Durant les années 1970, Kish a contribué à l'avancement dans ce domaine à titre de directeur de thèse de trois étudiants de doctorat de l'Université du Michigan (Erickson 1973; Kalsbeek 1973; Purcell et Kish 1979, 1980).

Parfois, il est possible de produire des estimations régionales directes ou fondées sur des données d'enquête. Les données du Recensement de la population sont une source manifeste d'estimations pour les domaines de toutes tailles et sont effectivement l'une des sources principales d'estimations régionales. Cependant, les données d'un recensement décennal deviennent périmées à mesure que l'on progresse vers la décennie suivante. Pour régler ce problème, Kish a proposé de remplacer le recensement par des échantillons à renouvellement complet, ou échantillons successifs, de sorte que l'on puisse produire des estimations plus à jour grâce à l'établissement de la collecte des données au fil du temps. Il a proposé une méthode de ce genre pour la première fois en 1979 (Kish 1979a,b) et, après cela, a rédigé de nombreux articles traitant de ce sujet (Kish 1981, 1983, 1986, 1990, 1997, 1998, 2002; Kish et Verma 1986). y compris celui sur la façon de cumuler les données d'enquête au fil du temps (Kish 1999). Dans un autre article du présent volume, Charles Alexander (2001) fait un compte rendu détaillé des travaux de Kish dans ce domaine et du rôle qu'il a joué dans l'élaboration de l'American Community Survey, enquête continue à grande échelle que le U.S. Census Bureau prévoit lancer pour remplacer le questionnaire détaillé du Recensement de 2010.

Problèmes spéciaux d'échantillonnage. Au cours de sa carrière, Kish a observé à maintes reprises des problèmes particuliers d'échantillonnage qui se manifestent fréquemment et a proposé certaines solutions efficaces. Les domaines auxquels il a contribué sont :

- **Echantillonnage de populations rares ou imprécises.** La production d'un plan d'échantillonnage efficace pour une population rare ou imprécise (comme les personnes atteintes d'une maladie rare ou les sans-abri) est l'un des plus grands défis que doivent relever les statisticiens d'enquêtes. Kish (1965b, 1991) a rédigé des revues instructives des méthodes qui permettent de s'attaquer à ce genre de problème.
- **Maximisation du chevauchement.** Lorsqu'une population est échantillonnée de façon répétée au fil du temps, il faut savoir comment contrôler le chevauchement entre les échantillons des cycles successifs. Un cas particulier est celui où l'on utilise un échantillon matriciel d'UPE et que celui-ci doit être mis à jour lorsque les nouvelles données de recensement sont diffusées. Fréquemment, il est souhaitable de maximiser le chevauchement dans l'échantillon de l'UPE tout en mettant à jour les mesures de taille et en modifiant la stratification afin de refléter les données courantes. Kish et Scott (1971) ont proposé une méthode assez simple et efficace pour répondre à ces exigences.
- **Echantillonnage d'organismes de taille inégale.** Certaines enquêtes sont conçues pour produire des estimations sur des unités de niveaux différents,

considérablement l'évolution des pratiques d'échantillonnage en particulier et des études par sondage en général. Les paragraphes qui suivent, qui donnent un aperçu de ses travaux, sont classés par sujet.

Estimation de la variance. Avant les années 1970, l'analyse des données d'enquête était très limitée, compte tenu des outils analytiques existants, c'est-à-dire principalement des machines à cartes perforées, telles que des compenseuses-trieuses et des totalisatrices, ainsi que des calculatrices manuelles. Donc, des statistiques comme les coefficients de pondération, particulièrement ceux qui n'étaient pas des nombres entiers, étaient difficiles à manipuler. C'est pour cette raison que Kish a étudié l'utilisation de coefficients de pondération uniformes grâce à la table de sélection de Kish, même si l'estimation non biaisée demandait l'application de poids proportionnels au nombre de membres admissibles du ménage.

À cause de la complexité des calculs, avant les années 1970, on estimait rarement les erreurs d'échantillonnage d'une façon qui reflétait les plans d'échantillonnage complexes habituellement employés pour les études par sondage. Une pratique courante consistait à calculer les variances comme si l'on avait sélectionné un échantillon aléatoire simple (EAS). Kish s'est efforcé de faire valoir l'utilisation des méthodes appropriées d'estimation de la variance auprès des chercheurs du ndp du maine social en illustrant la sous-estimation importante qui a souvent lieu lorsque l'on applique des formules d'EAS à des échantillons en grappes (Kish 1957). Au départ, il a mis au point et appliqué des méthodes de calcul simples, en insistant sur la simplicité d'un plan de sondage apparié en vertu duquel deux UPE sont échantillonnés dans chaque strate (Kish et Hess 1959a; Kish 1968). Il a inventé l'expression « effet de plan de sondage » pour désigner le rapport de la variance d'une estimation d'enquête obtenue au moyen d'un plan de sondage donné à la variance de la même estimation obtenue d'après un échantillon aléatoire simple de même taille. Il a utilisé abondamment ce concept dans son fameux traité intitulé *Survey Sampling* (Kish 1965a), qui décrit de façon encyclopédique la pratique de l'échantillonnage d'enquête et qui demeurera l'un des classiques publiés par Wiley il très fréquemment. Tout au long de sa carrière, Kish n'a cessé de s'intéresser aux effets de plan de sondage qu'il considérait comme un instrument important de conception et d'analyse des échantillons d'enquête (voir, par exemple, Kish 1982, 1995a; Kish, Frankel, Verma et Kaciroti 1995; Kish, Groves et Krotki 1976). Dans le cas d'un échantillon en grappes, un terme important de la formule de l'effet de plan de sondage est celui de la corrélation intraclass, dont Kish a discuté dans sa thèse de doctorat (Kish 1952) et dans plusieurs autres articles (par exemple, Kish 1954, 1961a).

Lorsque l'ordinateur a été développé, Kish a rapidement saisi l'importance de cet outil pour l'estimation de la variance et, avec ses collègues du SRC, a développé un premier logiciel intitulé *Sampling Error Program Package* (Kish, Frankel et Van Eck 1972). En collaboration avec

L'un de ces problèmes consistait à déterminer comment un intervieweur pouvait sélectionner aléatoirement une personne dans un ménage échantillonné. À l'époque, on avait mis au point des méthodes d'échantillonnage probabiliste pour l'échantillonnage des ménages que l'on appliquait à la Current Population Survey. Mais cette enquête était conçue pour recueillir des données sur tous les membres des ménages échantillonnés, si bien qu'il ne fallait sélectionner personne dans les ménages. Kish a inventé une méthode de sélection objective d'un répondant et l'a présentée par écrit dans un mémoire. Son collègue Angus Campbell a insisté pour qu'il soumette son travail à une revue pour publication, ce qui a donné le fameux article qui a été sa première étude publiée (Kish 1949). La méthode, dont l'usage est maintenant très répandu, porte le nom de table de sélection de Kish.

Le deuxième problème isolé par Kish était celui du dénombrement de la non-réponse. Pour défendre son idée du dénombrement et de la déclaration de la non-réponse dans le cas de l'échantillonnage probabiliste, il a dû argumenter contre ses collègues qui craignaient que cette pratique désavantage le SRC, relativement aux organismes utilisant des méthodes non probabilistes. Kish a eu gain de cause et le SRC a adopté sa méthode, qui est maintenant pleinement reconnue comme une norme de bonne pratique.

Le troisième problème était celui de la stratification en multiple. La stratification classique se fonde sur l'hypothèse de l'indépendance des sélections entre strates, le nombre maximal possible de strates étant égal au nombre de sélections. Surtout si le nombre de sélections est faible, comme cela est souvent le cas pour les unités primaires d'échantillonnage (UPE) dans un plan de sondage à plusieurs degrés, il est parfois souhaitable d'obtenir un échantillon mieux équilibré que ne le permet la stratification classique. En collaboration avec Roe Goodman, Kish a mis au point la technique de sélection contrôlée qui donne ce meilleur équilibre grâce à l'abandon de l'hypothèse de l'indépendance des sélections entre strates, tout en maintenant un échantillonnage probabiliste (Goodman et Kish 1950). Kish, qui cherchait toujours à trouver de bons noms, préférait qualifier la technique de « stratification multiple », expression qu'il a utilisée dans son traité sur l'échantillonnage (Kish 1965a).

Les travaux de recherche subséquents de Kish dans le domaine de la statistique d'enquête ont été de grande portée, couvrant de nombreux aspects des méthodes d'échantillonnage, des erreurs non dues à l'échantillonnage, de l'estimation régionale, de la conception d'enquêtes à composantes temporelle et spatiale et des études par observation. Ses nombreuses réalisations ont influencé

L'influence de Leslie Kish sur la statistique d'enquête

GRAHAM KALTON¹

RÉSUMÉ

Leslie Kish, l'un des pionniers de l'échantillonnage, est décédé le 7 octobre 2000 à l'âge de 90 ans. Le présent article donne un aperçu de l'influence qu'il a exercée sur la statistique d'enquête, principalement de ses travaux de recherche, mais aussi des efforts qu'il a déployés pour promouvoir l'application de méthodes d'échantillonnage probabilistes valides partout dans le monde. De grande portée, les travaux de recherche de Kish couvrent les méthodes d'échantillonnage, l'estimation de la variance et des effets de plans de sondage, les erreurs non dues à l'échantillonnage, l'estimation régionale, la conception d'enquête à dimensions temporelle et spatiale et les études par observation. Il a fait connaître les plans d'échantillonnage probabiliste grâce à ses activités d'expert-conseil dans de nombreux pays, à ses écrits et, tout spécialement, au programme d'étés efficaces sur l'échantillonnage à l'intention des statisticiens étrangers qu'il a créé au Survey Research Center de l'Université du Michigan.

MOTS CLÉS : Plan d'échantillonnage; estimation de la variance; erreurs non dues à l'échantillonnage; échantillons réussis.

1. INTRODUCTION

Leslie Kish, l'un des pionniers de l'échantillonnage, est décédé le 7 octobre 2000 à l'âge de 90 ans. Durant sa longue et fructueuse carrière, il a exercé une influence considérable dans le domaine, grâce à ses travaux de recherche impressionnants ainsi qu'à la manière extrêmement efficace avec laquelle il a fait connaître les méthodes scientifiques d'échantillonnage probabiliste partout dans le monde, surtout dans les pays en voie de développement. Ses travaux, dont la portée est vaste, se concentraient toujours sur des questions d'importance pratique et ses innovations ont facilité l'utilisation de méthodes efficaces d'échantillonnage probabiliste dans divers domaines. Il a mis en valeur la pratique de l'échantillonnage probabiliste grâce à ses écrits magistraux (destinés surtout aux sociologues et aux démographes), à ses nombreuses services de consultation et à la formation qu'il prodiguait aux statisticiens d'enquête, particulièrement ceux venant des pays en voie de développement.

Le présent article donne une idée de l'influence exercée par Kish sur la statistique d'enquête, surtout du rôle qu'il a joué dans l'avancement des méthodes d'échantillonnage d'enquête et, de façon plus générale, de ses travaux de recherche. Une brève description de sa carrière permettra de placer ses réalisations dans leur contexte. Les lecteurs qui souhaiteraient avoir plus de détails sur la vie fascinante de Kish sont invités à consulter le compte rendu de son entretien de 1994 avec Frankel et King (1996). Certains renseignements présentés ici en sont extraits.

Kish est né en 1910 à Poprad, qui faisait alors partie de l'empire Austro-hongrois, mais qui appartenait aujourd'hui à la Slovaquie. En 1926, lui et sa famille ont immigré aux États-Unis. À la suite du décès de son père l'année suivante,

2. RECHERCHE

Il a occupé un poste d'assistant de laboratoire au Rockefeller Institute for Medical Research, alors qu'il fréquentait la Bay Ridge Evening High School. Il a obtenu son diplôme d'études secondaires en 1930 et s'est inscrit au cours du soir du College of the City of New York tout en continuant à travailler 54 heures par semaine au Rockefeller Institute. C'est durant son travail à l'Institut qu'il s'est passionné pour la statistique et qu'il a commencé à étudier de lui-même les traités rédigés par Fisher, Yule, Wallace et Snedecor, Tippett, Pearl et d'autres. En 1937, il a interrompu ses études pour se joindre à l'International Brigade dans sa lutte pour la cause loyaliste durant la guerre civile d'Espagne. Il est retourné aux États-Unis en 1939 où il a obtenu un baccalauréat en mathématiques avec distinction la même année. Après avoir été recruté par le U.S. Census Bureau à titre de chef de section, il a obtenu un poste de statisticien à la Division of Program Surveys du United States Department of Agriculture (USDA). En 1942, il a quitté la Division parce qu'il était appelé sous les drapeaux et y est retourné en 1945 une fois la guerre terminée. En 1947, il a déménagé avec un groupe de collègues de l'USDA dirigé par Rensis Likert afin d'établir le Survey Research Center à l'Université du Michigan, où il est resté jusqu'à sa retraite en 1981, après être devenu professeur-émérite. Il a poursuivi activement ses activités professionnelles jusqu'à son décès en 2000.

Au début de la carrière de Kish, la science de l'échantillonnage était à ses premiers balbutiements. Nombre d'enquêtes par sondage se fondaient sur des échantillons non probabilistes. Les méthodes d'échantillonnage

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

- TALLIS, G.M. (1978). Note on robust estimation in finite populations. *Sankhyā C*, 40, 136-138.
- TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.
- THEBERGE, A. (2000). Calage et poids restreints. *Techniques d'enquête*, 26, 113-122.
- TITLE, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.
- TREMBLAY, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête*, 12, 91-103.
- WATSON, D. J. (1937). The estimation of leaf area in field crops. *Journal of Agricultural Science*, 27, 474-483.
- WOODRUFF, R.S., et CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.
- WU, C., et SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- YATES, F. (1949). *Sampling Methods for Census and Surveys*. London: Griffin.
- YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.
- ZYSKIND, G. (1976). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38, 1092-1109.

- ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- ROYALT, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALT, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Association*, 71, 657-664.
- ROYALT, R.M. (1986). The prediction approach to robust variance estimation in two stage cluster sampling. *Journal of the American Statistical Association*, 81, 119-123.
- ROYALT, R.M., et CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- ROYALT, R.M., et CUMBERLAND, W.G. (1981). The finite population linear regression estimator and estimators of its variance, an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRDAL, C.-E. (1980). On π -weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- SÄRDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistics Association*, 91, 1289-1300.
- SÄRDAL, C.-E., SWENSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SÄRDAL, C.-E., SWENSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SÄRDAL, C.-E., et WRIGHT, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- SCOTT, A., et SMITH, T.M.F. (1974). Linear superpopulation models in survey and sampling. *Sankhyā*, C, 36, 143-146.
- SCOTT, A., et WU, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.
- SILVA, P.L.D.N., et SKINNER, C.J. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.
- SINGH, A.C., et FOLSOM, R.E. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 610-615.
- SINGH, A.C., KENNEDY, B., et WU, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.
- SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- KUO, L. (1988). Classical and prediction approaches to estimating distribution function from survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 280-285.
- LAZZERONI, L.C., et LITTLE, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- MADOW, W.G., et MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.
- MICKEL, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- MONTANARI, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.
- MONTANARI, G.E. (1999). A study on the conditional properties of finite population mean estimators. *Metron*, 57, 21-35.
- MUKHOPADHAYAY, P. (1993). Estimation of a finite population total under regression models: A review. *Sankhyā*, 55, 141-155.
- NIJEWENBROEK, N., RENNESSEN, R., et HOFMAN, L. (2000). Towards a generalized weighting system. Dans *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia.
- PARK, M. (2002). Regression estimation of the mean in Survey Sampling. Thèse de doctorat non-publiée dissertation, Iowa State University, Ames, Iowa.
- PFEFFERMAN, D. (1984). Note on large sample properties of balanced samples. *Journal of the Royal Statistical Society, Series B*, 46, 38-41.
- RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- RAO, J.N.K. (2002). *Small Area Estimation Theory and Methods*. New York: John Wiley & Sons, Inc.
- RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-491.
- RAO, J.N.K., et SINGH, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.
- ROBINSON, G.K. (1991). The BLUE is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-32.
- ROBINSON, P.M., et SÄRDAL, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā, Series B*, 45, 240-248.

- FOLSOM, R.E., et SINGH, A.C. (2000). The generalized exponential model for a unified approach to sampling weight calibration and outlier weight treatment, nonresponse adjustment and post-stratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 598-603.
- FRANKEL, M.R. (1971). Inference from survey samples: An empirical investigation. Institute for Social Research, University of Michigan, Ann Arbor.
- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the America Statistical Association*, 61, 1172-1183.
- FULLER, W.A. (1973). Regression for sample surveys. Un article présenté à la réunion de International Statistical Institute, Août, 1973, Vienna, Austria.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhya Series C*, 37, 117-132.
- FULLER, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-137.
- FULLER, W.A., et AN, A.B. (1998). Regression adjustments for nonresponse. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.
- FULLER, W.A., et BATTESE, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.
- FULLER, W.A., et ISAKI, C.T. (1981). Survey design under superpopulation models. Dans *Current Topics in Survey Sampling*, (D. Krewski, J. N. K. Rao et R. Plalek, Eds.), New York: Academic Press, 199-226.
- FULLER, W.A., LOUGHIN, M.M. et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20 79-89.
- FULLER, W.A., et RAO, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.
- GAMBINO, J., KENNEDY, B. et SINGH, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 69-79.
- GEROW, K., et MCCULLOCH, C.E. (2000). Simultaneously model unbiased, design-unbiased estimation. *Biometrics* 56, 873-878.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.
- GODAMBE, V.P., et JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.
- GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369-375.
- GRAYBILL, F.A. (1976). *Theory and application of the linear model*. Wadsworth, Belmont, CA.
- HÄYER, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Pestovani Matematiky*, 84, 387-423.
- KALTON, G., et MALIGALIG, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference*, U. S. Bureau of the Census, 409-428.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples (avec discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KONUN, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association*, 57, 590-606.
- KOTT, P.S. (1994). A note on handling nonresponses in sample surveys. *Journal of the American Statistical Association*, 693-696.
- HANURAV, T.V. (1966). Some aspects of unified sampling theory. *Sankhya, Series A*, 28, 175-204.
- HENDERSON, C.R. (1963). Selection index and expected genetic advance. Dans *Statistical Genetics and Plant Breeding*, 141-163. Publication 982, Washington, DC.
- HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include estimation of random effects. *Annals of Statistics*, 4, 384-395.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., FULLER, W.A. et HICKMAN, R.D. (1978). *Super Cap*, (6^e édition, 1980) Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., SÄRNDALE, C.-E. et BINDER, D. A. (1995). Weighting and estimation in business surveys. *Business Survey Methods*, (Eds. Cox, Binder, Chinappa, Colledge et Kott) New York: John Wiley & Sons, Inc., 477-502.
- HOLT, D. et SMITH, T.M. F. (1979). Post Stratification. *Journal of the Royal Statistical Society, Serie. A*, 142, 33-46.
- HORN, S.D., HORN, R.A. et DUNCAN, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the social statistics section, American Statistical Association*, 300-305.
- HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.
- ISAKI, C.T. (1970). Survey designs utilizing prior information. Thèse de doctorat non-publiée. Iowa State University.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- ISAKI, C.T., TSAY, J.H. et FULLER, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 26, 37-49.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agriculture Experiment Station Research Bulletin*, 304
- KALTON, G., et MALIGALIG, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference*, U. S. Bureau of the Census, 409-428.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples (avec discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KONUN, H.S. (1962). Regression analysis for sample surveys. *Journal of the American Statistical Association*, 57, 590-606.
- KOTT, P.S. (1994). A note on handling nonresponses in sample surveys. *Journal of the American Statistical Association*, 693-696.

BIBLIOGRAPHIE

- ANDERSON, C., et NORDBERG, L. (1998). A user's guide to CLAN97. Statistics Sweden, Örebro, Sweden.
- BANKIER, M.D., RATHWELL, S. et MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, Direction de la méthodologie, Census Operations Section, Social Survey Methods Division, Statistique Canada, Ottawa.
- BANKIER, M.D., HOULE, A.M. et LUC, M. (1997). Calibration estimation in the 1991 and 1996 Canadian census. Statistique Canada (draft), 8 pages.
- BARDSELY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- BETHLEHEM, J.G. et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- BISHOP, Y.M.M., FIENBERG, S.E. et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- BREIDT, F.J., et OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.
- BREWER, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- BREWER, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- BREWER, K.R.W., HANIF, M. et TAM, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.
- BRICK, J.M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions de service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.
- CASSEL, C.M., SÄRNDAAL, C.-E. et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized design-based principles for finite populations when model-based and prediction theory for finite populations are combined. *Scandinavian Journal of Statistics*, 6, 97-106.
- CASSEL, C.M., SÄRNDAAL, C.-E. et WRETMAN, J.H. (1979). Prediction theory for finite populations when model-based and design-based principles are combined. *Scandinavian Journal of Statistics*, 3, 143-160.
- CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- CHAMBERS, R.L., DORMAN, A.H. et WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3^e éd., New York: John Wiley & Sons, Inc.
- COOK, R.D., et WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEMING, W.E., et STEPHAN, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 45-49.
- DEVILLE, J., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J., SÄRNDAAL, C.-E. et SAVTOR, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DORMAN, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.
- DORMAN, A.H., et HALL, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1475.
- DUCHESNE, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16, 133-138.
- DU MOUCHEL, W. H., et DUNCAN, G. J. (1983). Using survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- ELTINGE, J.L., et YANSANEH, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.
- ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.
- ESTEVAO, V., HIDIROGLOU, M.A. et SÄRNDAAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- FIRTH, D., et BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.
- FOLSON, R.E., et WITT, M.B. (1994). Testing a new attention nonresponse adjustment method for SIPP. *Proceedings of the Association, 428-433*.
- Fuller : Estimation par régression appliquée à l'échantillonnage

$$V\{\bar{\mathbf{h}}_{HT}\} = O^p(n^{-1}),$$

$$V\{\bar{\mathbf{b}}_{HT}\} = V\{\bar{\mathbf{b}}_{HT}\} + O^p(n^{-1})$$

car $\delta_p = O^p(n^{-1/2})$. Le résultat (A.7) découle alors de la normalité asymptotique de $\hat{\beta} - \beta_N$.

Théorème A.2. Soit $\mathbf{y}' = (y_1, y_2, \dots, y_n)$ et $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \dots, \mathbf{x}_n')$. Supposons que est une matrice $n \times n$ symétrique non singulière et que Φ_N est une matrice $N \times N$ symétrique non singulière. Supposons que

$$\bar{\mathbf{y}}_N, \bar{\mathbf{x}}_N, n^{-1}(\mathbf{X}'\Phi^{-1}\mathbf{X}) \text{ and } n^{-1}\mathbf{X}'\Phi^{-1}\mathbf{y}$$

sont des estimateurs convergents par rapport au plan d'échantillonnage des caractéristiques de la population finie $\bar{\mathbf{y}}_N, \bar{\mathbf{x}}_N, \mathbf{O}^{xxN}$ et \mathbf{O}^{xyN} , respectivement, où

$$[\mathbf{O}^{xxN}, \mathbf{O}^{xyN}] = [N^{-1}\mathbf{X}'\Phi^{-1}\mathbf{X}, N^{-1}\mathbf{X}'\Phi^{-1}\mathbf{y}] \quad (\text{A.9})$$

Soit $\beta_N = \mathbf{O}^{-1} \mathbf{O}^{xxN}$. Supposons qu'il existe une suite de vecteurs colonnes $\{\gamma_N\}$ tels que

$$\mathbf{X} \gamma_N = \Phi \mathbf{D}_N^{-1} \mathbf{y}_N$$

pour tous les échantillons possibles, où $\mathbf{D}_N = \text{diag}(\pi_1, \pi_2, \dots, \pi_n)$ et \mathbf{y}_N est un vecteur colonne à n dimensions de 1. Alors, l'estimateur par régression $\bar{\mathbf{x}}_N \hat{\beta}$ pour lequel

$$\hat{\beta} = (\mathbf{X}'\Phi^{-1}\mathbf{X})^{-1}\mathbf{X}'\Phi^{-1}\mathbf{y}_N, \quad (\text{A.11})$$

est un estimateur convergent par rapport au plan d'échantillonnage de $\bar{\mathbf{x}}_N$.

Preuve. Si $\hat{\beta}$ est défini par (A.11), alors, compte tenu des propriétés des estimateurs par les moindres carrés généralisés,

$$(\mathbf{y} - \mathbf{X}\hat{\beta})'\Phi^{-1}\mathbf{X} = \mathbf{0},$$

Si la condition (A.10) est satisfaite, alors

$$(\mathbf{y} - \mathbf{X}\hat{\beta})'\mathbf{D}_N^{-1}\mathbf{I} = \left(\sum_{i \in A} \pi_i^{-1} \right) (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_N \hat{\beta}) = \mathbf{0}.$$

Il s'ensuit que $\bar{\mathbf{y}}_{reg}$ est convergent par rapport au plan d'échantillonnage, car

$$0 = p \lim_{N \rightarrow \infty} \left\{ (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_N \hat{\beta}_N)' | \mathbf{F}_N \right\}$$

$$= p \lim_{N \rightarrow \infty} \left\{ (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_N \hat{\beta}_N)' | \mathbf{F}_N \right\}$$

$$= p \lim_{N \rightarrow \infty} \left\{ (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_N \hat{\beta}_N)' | \mathbf{F}_N \right\}.$$

Théorème A.3. Soit une suite de populations et d'échantillons tels que définis dans le théorème A.1.

Supposons que \mathbf{z}_i est un vecteur de la forme $\mathbf{z}_i = (y_i, 1, \mathbf{x}_{1,i})'$ et que $\mathbf{z}_{1,N} = (y_N, 1, \mathbf{x}_{1,N})'$. Supposons que $\bar{\mathbf{z}}_{1,N}$ est un estimateur convergent par rapport au plan d'échantillonnage de la moyenne par rapport au plan d'échantillonnage des covariances non singulières

$$V\{\bar{\mathbf{z}}_{1,N} | \mathbf{F}_N\} = O(n^{-1}) \quad (\text{A.12})$$

$$n^{1/2}(\bar{\mathbf{z}}_{1,N} - \bar{\mathbf{z}}_{1,N}) | \mathbf{F}_N \xrightarrow{L} N(\mathbf{0}, \Sigma^{zz}), \quad (\text{A.13})$$

où Σ^{zz} est la limite de $n V\{\bar{\mathbf{z}}_{1,N} | \mathbf{F}_N\}$. Supposons qu'il existe un estimateur de la variance de $\bar{\mathbf{z}}_{1,N}$, noté $V\{\bar{\mathbf{z}}_{1,N}\}$, tel que

$$p \lim_{N \rightarrow \infty} n^{1/2} V\{\bar{\mathbf{z}}_{1,N} | \mathbf{F}_N\} - V\{\bar{\mathbf{z}}_{1,N} | \mathbf{F}_N\} = 0 \quad (\text{A.14})$$

pour $\delta > 0$. Représentons par $\hat{\beta}_{1,dopt}$ le vecteur qui minimise

$$V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\} \quad (\text{A.15})$$

et par $\hat{\beta}_{1,dopt}$ le vecteur qui minimise $V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\}$. Supposons que $\bar{\mathbf{y}}_{d,reg}$ est défini par (4.29). Alors, $\bar{\mathbf{y}}_{d,reg}$ est, parmi les estimateurs convergents par rapport au plan d'échantillonnage de la forme $\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}$ celui dont la variance limite est minimale. En outre,

$$\bar{\mathbf{y}}_{d,reg} - \bar{\mathbf{y}}_N \xrightarrow{L} N(0, 1), \quad (\text{A.16})$$

où $V\{\bar{\mathbf{e}}_N\}$ est l'estimateur (A.14) construit avec $\bar{\mathbf{e}}_i = y_i - \bar{y}_N - (\bar{\mathbf{x}}_{1,i} - \bar{\mathbf{x}}_{1,N})' \hat{\beta}_{1,dopt}$

$$\hat{\beta}_{1,dopt} = \left[V\{\bar{\mathbf{x}}_{1,N}\}^{-1} C\{\bar{\mathbf{x}}_{1,N}, \bar{\mathbf{y}}_N\} \right]$$

minimise l'estimation de la variance de (A.15) et, en vertu de l'hypothèse (A.14), l'estimation de la variance converge vers la variance réelle. Donc, $\hat{\beta}_{1,dopt}$ est convergent par rapport au plan d'échantillonnage pour $\hat{\beta}_{1,dopt}$ et $\hat{\beta}_{1,dopt}$ minimise $V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\}$. Par conséquent, aucun estimateur de la forme (4.29) n'a une distribution limite dont la variance est plus faible.

Maintenant

$$\bar{\mathbf{y}}_{d,reg} - \bar{\mathbf{y}}_N = \bar{\mathbf{y}}_N - \bar{\mathbf{y}}_N - (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,N})' \hat{\beta}_{1,dopt} = \bar{\mathbf{e}}_N + O^p(n^{-1/2}),$$

où $\bar{\mathbf{e}}_i = y_i - \bar{y}_N - (\bar{\mathbf{x}}_{1,i} - \bar{\mathbf{x}}_{1,N})' \hat{\beta}_{1,dopt}$. Par conséquent, la variance de la distribution limite de $n^{1/2}(\bar{\mathbf{y}}_{d,reg} - \bar{\mathbf{y}}_N)$ est la variance de $n^{1/2}(\bar{\mathbf{e}}_N - \bar{\mathbf{e}}_N)$. D'après l'hypothèse (A.14), l'estimateur $V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\}$ pour toute valeur fixée de γ . Puisque la variance $V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\}$ converge de la variance $V\{\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \hat{\beta}_{1,dopt}\}$, l'estimation de la variance fondée sur $\bar{\mathbf{e}}_i$ converge vers l'estimation de la variance fondée sur $\bar{\mathbf{e}}_i$ et l'expression (A.16) est vérifiée.

ANNEXE

$$(A.6) \quad n \left(\hat{V}^{zz} - V^{zz} \mid F^N = O^p(n^{-1/3}) \right) \quad \text{pour tout } z \text{ dont le quatrième moment est borné. Alors}$$

$$(A.7) \quad [V\{\hat{\beta}\}]^{-1/2} [\hat{\beta} - \beta_N] \mid F^N \xrightarrow{L} N(0, I),$$

$$(A.8) \quad V\{\hat{\beta}\} = \hat{Q}^{-1}_{xx} \hat{V}^{bb} \hat{Q}^{-1}_{xx}, \quad \text{où}$$

$\hat{V}^{bb} = V\{\hat{b}^{HT}\}$ est l'estimation de la variance liée au plan d'échantillonnage de \hat{b}^{HT} calculée avec $\hat{b}^i_i = n^{-1} N \pi_i \zeta_i^i e_i^i$ et $\hat{e}^i_i = y_i^i - \hat{x}_i^i \hat{\beta}_i$. Preuve. L'erreur dans $\hat{\beta}$ est

$$\hat{\beta} - \beta_N = (X' \Phi^{-1} X)^{-1} [X' \Phi^{-1} Y - X' \Phi^{-1} X \beta_N]$$

$$= \hat{Q}^{-1}_{xx} (n^{-1} N \pi_i \Phi^{-1} e_i).$$

Maintenant, $\hat{\beta}$ est un estimateur par les moindres carrés généralisés. Par conséquent,

$$\hat{e} \Phi^{-1} X = (Y - X \hat{\beta})' \Phi^{-1} X = 0$$

$$(A.1) \quad \text{et } \hat{Q}^{xyN} = \hat{\beta}'^N \hat{Q}^{xxN} = \hat{Q}^{xxN} = 0. \text{ Selon l'hypothèse (A.1)}$$

$$\hat{Q}^{xx} = n^{-1} N \pi_i \Phi^{-1} e_i = O^p(n^{-1/2}).$$

Donc,

$$\hat{\beta} - \beta_N = \hat{Q}^{-1}_{xx} \left(n^{-1} \sum_{i \in A} \zeta_i^i e_i^i + O^p(n^{-1}) \right)$$

$$= \hat{Q}^{-1}_{xxN} \left(N^{-1} \sum_{i \in A} \pi_i^i b^i_i + O^p(n^{-1}) \right).$$

Les quatrième moments des b^i_i sont bornés par les hypothèses. Donc, selon l'hypothèse (A.5)

$$V^{bb}_{1/2} (\hat{\beta} - \beta_N) \xrightarrow{L} N(0, I),$$

où

$$V^{bb} = \hat{Q}^{-1}_{xx} V^{bb} \hat{Q}^{-1}_{xxN}$$

et $V^{bb} = V\{\hat{b}^{HT}\}$. Maintenant,

$$n^{-1} N \pi_i \Phi^{-1} e_i = n^{-1} N \pi_i \Phi^{-1} X' \Phi^{-1} X' \Phi^{-1} e_i + n^{-1} N \pi_i \Phi^{-1} X' \Phi^{-1} X' \Phi^{-1} e_i$$

$$=: N^{-1} \sum_{i \in A} \pi_i^i b^i_i + N^{-1} \sum_{i \in A} \pi_i^i h^i_i,$$

où

$$h^i_i = n^{-1} N \pi_i \zeta_i^i x_i^i \delta^i_i$$

et $\delta^i_i = \beta_N - \hat{\beta}$. Pour toute valeur fixée de δ , selon (A.6), l'estimation de la variance de $N^{-1} \sum_{i \in A} \pi_i^i (b^i_i + h^i_i)$ converge vers la variance de l'estimateur de la moyenne de $b + h$. Par hypothèse, les éléments de $\zeta_i^i x_i^i$ ont un quatrième moment. Pour une valeur fixée de δ , la variance de h^{HT} est $O(n^{-1})$. Pour $\delta = \delta^p$

Cette annexe contient les théorèmes qui appuient les propriétés limites des estimateurs par régression décrits à la section 4.

Théorème A.1. Supposons que $\{U^N, F^N, A^N, n^N; N = k+3, k+4, \dots\}$ est une suite de populations finies et d'échantillons, où F^N est un échantillon provenant d'une population infinie ayant huit moments, A^N est l'échantillon de taille n^N sélectionné à partir de la $N^{\text{ème}}$ population. Supposons que β est défini par (4.4) et que

$$\hat{Q}^{zz} = n^{-1} Z' \Phi^{-1} Z,$$

où Φ est une matrice symétrique définie positive qui peut être une fonction de X , mais non de Y . Z est défini conformément à (4.2), et nous omettons l'indice N sur les quantités d'échantillon. Supposons que \hat{Q}^{zz} est définie positive et que sa probabilité est égale à un. Si Φ est aléatoire, supposons que les lignes de $\Phi^{-1} Z$ ont borné le quatrième moment. Supposons que les probabilités de sélection satisfont

$$0 < K_1 < N n^{-1} \pi_i < K_2,$$

[$(\hat{z}^{HT} - \bar{z}^N)' (\hat{Q}^{zz} - Q^{zzN}) \mid F^N = O^p(n^{-1/3})$] (A.1) où $\hat{z}^{HT} = (\hat{y}^{HT}, \bar{x}^{HT})' = N^{-1} \sum_{i \in A} \pi_i^i z^i_i$ (A.2) $Q^{zzN} = E\{\hat{Q}^{zz} \mid F^N, \bar{z}^N\}$ est la moyenne de la population finie de z , Q^{zzN} est une matrice définie positive pour la $N^{\text{ème}}$ population et Q^{zz} est définie et positive. Alors,

$$(A.3) \quad \hat{\beta} - \beta_N \mid F^N = O^{-1}_{xxN} \bar{b}^{HT} + O^p(n^{-1}),$$

$$\text{où } \hat{\beta}^N = O^{-1}_{xxN} \bar{b}^{HT} = N^{-1} \sum_{i \in A} \pi_i^i b^i_i, b^i_i = n^{-1} N \pi_i \zeta_i^i e_i^i,$$

$$Q^{zzN} = \begin{pmatrix} Q^{xxN} & Q^{xyN} \\ Q^{yxN} & Q^{yyN} \end{pmatrix},$$

et $\zeta_i^i b^i_i = y_i^i - x_i^i \beta^N$, ζ_i^i est la colonne i de $X' \Phi^{-1}$.

Supposons que le plan d'échantillonnage est tel que

$$(A.5) \quad V^{zz}_{1/2} \{ \bar{z}^{HT} - \bar{z}^N \mid F^N \} \xrightarrow{L} N(0, I),$$

lorsque $n^N \rightarrow \infty$ pour tout z dont le quatrième moment est fini, où V^{zz} est la matrice des covariances de $\bar{z}^{HT} - \bar{z}^N$. Supposons que V^{zz} est $O(n^{-1})$ et que le plan d'échantillonnage admet un estimateur \hat{V}^{zz} tel que

(1993) discutent de l'extension de la méthode itérative du quotient à des variables x générales et à des extensions visant à inclure les bornes sur les coefficients de pondération.

Tillé (1998) a proposé d'utiliser des probabilités conditionnelles approximatifs, contraintes sur \mathbf{x}_i^u , pour calculer un estimateur. On peut étendre son approximation de façon à produire des coefficients de pondération par régression positifs avec une forte probabilité. Représentons par $\mathbf{x}_i^{(i)}$ un estimateur obtenu par suppression de l'élément i , ou de l'unité primaire d'échantillonnage i , et par modification des coefficients de pondération restants de sorte que $\mathbf{x}_i^{(i)}$ soit sans biais, ou converge vers le même ordre que \mathbf{x}_i^u pour la moyenne de la population de tous les éléments, sauf i . L'estimateur $\mathbf{x}_i^{(i)}$ peut être celui utilisé pour produire les écarts jackknife. Représentons par \mathbf{Z}_{xx}^u un estimateur de la matrice des covariances de \mathbf{x}_i^u et par $\mathbf{Z}_{xx}^{(i)}$, un estimateur conditionnel sur $i \in A$. Alors, dans les grands échantillons, \mathbf{x}_i^u et $\mathbf{x}_i^{(i)}$ suivent approximativement la loi de distribution normale et un estimateur de la probabilité que i soit incluse dans l'échantillon, étant donné l'estimateur de la moyenne \mathbf{x}_i^u , est

$$\pi_i^A = P\{i \in A \mid \mathbf{F}_i, \mathbf{x}_i^u\} \\ = \pi_i^t | \mathbf{Z}_{xx}^u |^{1/2} | \mathbf{Z}_{xx}^{(i)} |^{-1/2} \exp \{0.5 (\mathbf{G}_{xx}^u - \mathbf{G}_{xx}^{(i)})\} \quad (9.1)$$

où

$$\mathbf{G}_{xx}^u = (\mathbf{x}_i^u - \bar{\mathbf{x}}^u)(\mathbf{x}_i^u - \bar{\mathbf{x}}^u)', \\ \mathbf{G}_{xx}^{(i)} = (\mathbf{x}_i^{(i)} - \bar{\mathbf{x}}^{(i)})(\mathbf{x}_i^{(i)} - \bar{\mathbf{x}}^{(i)})',$$

et $\bar{\mathbf{x}}_i^u = (N - 1)^{-1} (\mathbf{N} \bar{\mathbf{x}}_i^u - \mathbf{x}_i^u)$. Dans le cas de l'échantillonnage aléatoire simple, Tillé (1998) a montré que l'estimateur

$$\bar{\mathbf{y}}^u = N^{-1} \sum_{i \in A} \mathbf{y}_i^u \quad (9.2)$$

où π_i^A est la probabilité conditionnelle calculée sous l'hypothèse de normalité, est approximativement égal à l'estimateur par régression. Comme l'estimateur n'est pas calé, nous proposons une version calée obtenue par calcul de l'estimateur par régression en prenant les π_i^A comme coefficients de pondération initiaux. La différence entre l'estimateur (9.2) et l'estimateur par régression construit en prenant les poids initiaux π_i^A est $O_p(n^{-1})$. Donc, il y a beaucoup de chances que les coefficients de pondération obtenus par régression conçus de la sorte soient positifs. L'estimateur de la variance $\mathbf{Z}_{xx}^{(i)}$ est assez simple à calculer pour les échantillons stratifiés, mais demande des calculs considérables dans d'autres cas. Donc, on pourrait choisir d'approximer $\mathbf{Z}_{xx}^{(i)}$. Étant donné que les coefficients de pondération obtenus par régression sont construits par minimisation de la fonction objective, on peut ajouter des contraintes au

problème afin d'imposer des bornes aux coefficients de pondération. Huang et Fuller (1978) ont proposé une méthode itérative équivalant à la construction, à chaque étape, d'une matrice Φ qui réduit le coefficient de pondération sur les observations pour lesquelles l'écart du coefficient de pondération courant par rapport à la moyenne est important en valeur absolue.

Afin de discuter de méthodes supplémentaires associées aux fonctions objectives quadratiques, supposons que nous ayons une matrice des covariances de travail, que nous représentons par Φ^e , pour le modèle (5.1) que nous utiliserons pour construire un estimateur par régression. Posons que \mathbf{a} est le vecteur colonne des coefficients de pondération initiaux et que $\Phi^e \mathbf{a}$ se trouve dans l'espace colonne de \mathbf{X} . Alors, les coefficients de pondération qui minimisent la variance conditionnelle due au modèle sont ceux qui minimisent $\mathbf{w}' \Phi^e \mathbf{w}$ ou, de façon équivalente, qui minimisent

$$(\mathbf{w} - \mathbf{a})' \Phi^e (\mathbf{w} - \mathbf{a}) \quad (9.3)$$

à condition que soit satisfaite la contrainte

$$\mathbf{w}' \mathbf{X} = \bar{\mathbf{y}}^u. \quad (9.4)$$

Étant donné une fonction objective, nous pouvons appliquer des contraintes supplémentaires aux \mathbf{w}_i de sorte que

$$L_1 \leq \mathbf{w}_i' \leq L_2, \quad i \in A, \quad (9.5)$$

où L_1 et L_2 sont des constantes non négatives. Minimiser (9.3), en veillant à ce que les contraintes (9.4) et (9.5) soient satisfaites, est un problème de programmation quadratique. L'utilisation de cette dernière a été proposée par Husain (1969) et utilisée par Isaki, Tsay et Fuller (2000).

Si le nombre de variables de contrôle utilisées est élevé, il peut être impossible de construire des coefficients de pondération qui satisfont les contraintes de calage et qui sont également compatibles entre des bornes raisonnables. Le praticien doit donc faire des compromis. La pratique la plus courante consiste à supprimer certaines variables du modèle. À cet égard, consultant Bankier, Rathwell et Majtkowski (1992) et Silva et Skinner (1997). Pour examiner une autre méthode, considérons la situation où certaines contraintes sont nécessaires, mais où d'autres peuvent être relâchées. Partitionnons la matrice des observations sur les variables auxiliaires de façon telle que

\mathbf{X}_0 (ou \mathbf{X}_2), où \mathbf{X}_0 représente l'ensemble de variables pour lequel des contraintes rigoureuses sont nécessaires et \mathbf{X}_2 ceux pour lequel les contraintes peuvent être relâchées. Supposons que $\Phi^e \mathbf{a}$ se situe dans l'espace colonne de \mathbf{X}_0 . Alors, une généralisation de (9.3) et (9.4) donne la fonction

$$(\mathbf{w} - \mathbf{a})' \Phi^e (\mathbf{w} - \mathbf{a}) + (\mathbf{w}' \mathbf{X}_2 - \bar{\mathbf{y}}_2^u)' \mathbf{W} (\mathbf{w}' \mathbf{X}_2 - \bar{\mathbf{y}}_2^u)' \quad (9.6)$$

et la contrainte

$$\mathbf{w}' \mathbf{X}_0 = \mathbf{x}_0^u = 0, \quad (9.7)$$

$$\gamma^N = \left(\sum_{i \in U} x_i' d_i' x_i' \right)^{-1} \sum_{i \in U} x_i' d_i' \gamma_i' \quad (8.1)$$

Nous pouvons exprimer la moyenne de population de y sous la forme

$$\bar{y}^N = \bar{x}^N \bar{\gamma}^N + \bar{a}^N \quad (8.2)$$

où $a_i' = y_i' - x_i' \gamma^N$ et \bar{a}^N est la moyenne de population des a_i' . L'estimateur par régression $\bar{y}_{reg}^N = \bar{x}^N \bar{\beta}$ sera convergent si la limite de la probabilité de \bar{a}^N est nulle. Cette limite sera nulle si la série de populations finies est une série d'échantillons aléatoires provenant d'une population infinie dans laquelle

$$\gamma_i' = \beta \gamma_i' + e_i' \quad (8.3)$$

et les e_i' de l'échantillon sont indépendantes de x_i' en ayant $E\{e_i' | x_i'\} = 0$.

Où bien, une condition suffisante pour que \bar{a}^N soit nulle est qu'il existe un vecteur colonne ξ tel que

$$x_i' \xi = d_i^{-1} \quad (8.4)$$

pour $i = 1, 2, \dots, N$. Donc, l'inverse de la probabilité de réponse est une fonction linéaire des variables de contrôle, l'estimateur par régression est un estimateur convergent de la moyenne de y . Un moyen de satisfaire (8.4) consiste à ce que les éléments de x_i' soient des variables nominales qui définissent des sous-groupes et que les probabilités de réponse soient constantes dans chaque sous-groupe.

Si (8.4) est satisfaite et que la probabilité de réponse est indépendante d'une unité à l'autre, alors l'estimateur de la variance fondée sur (4.12) est un estimateur approprié de la variance de l'estimateur par régression de la moyenne. Il est particulièrement important d'utiliser un estimateur de la variance de la forme (4.12) ou (4.25), et non de la forme (4.26), car $\bar{x}^N - \bar{x}^N$ n'est, en général, pas $O_p(n^{-1/2})$ de non-réponse. Singh et Folsom (2000) présentent un argument comparable pour l'estimateur de la variance (4.25) lorsque l'on utilise la régression pour faire la correction pour l'erreur de couverture.

Souvent, on procède à un rajustement préliminaire des probabilités de sélection pour tenir compte de la non-réponse, puis à l'estimation par régression. La méthode de rajustement de la réponse utilisée le plus fréquemment consiste à créer des cellules (strates a posteriori) et à ajuster par la méthode du ratio les coefficients de pondération calculés pour les répondants compris dans la cellule de sorte que la somme de ces coefficients soit égale au total estimé (ou connu) pour la cellule. Consulter, par exemple, Little et Rubin (1987, page 250). Les méthodes fondées sur une fonction estimée de probabilité de réponse sont examinées par Cassel, Särndal et Wretman (1983), Rosenbaum et Rubin (1983), Folsom et Witt (1994), Fuller et An (1998), et Folsom et Singh (2000). Brick, Waksberg et Keener (1996) utilisent une probabilité estimative de contact pour faire la correction pour tenir compte de la couverture de la base de sondage.

9. CONSIDÉRATIONS PRATIQUES

Si l'on veut utiliser les coefficients de pondération obtenus par régression dans le cas d'une enquête générale, aucun coefficient de pondération individuel n'utilise pour estimer un total ne devrait être inférieur à l'unité. En outre, il semble raisonnable, par souci de robustesse, d'éviter les coefficients de pondération ayant une valeur très grande. Nous examinons certaines méthodes mises au point afin de réaliser ces objectifs.

Plusieurs algorithmes produisent des coefficients de pondération positifs avec une probabilité élevée. Les méthodes itératives du quotient (raking ratio) produisent des coefficients de pondération positifs pour la plupart des configurations de données. Deville, Särndal et Sautory

Plusieurs algorithmes produisent des coefficients de pondération positifs avec une probabilité élevée. Les méthodes itératives du quotient (raking ratio) produisent des coefficients de pondération positifs pour la plupart des configurations de données. Deville, Särndal et Sautory

Plusieurs algorithmes produisent des coefficients de pondération positifs avec une probabilité élevée. Les méthodes itératives du quotient (raking ratio) produisent des coefficients de pondération positifs pour la plupart des configurations de données. Deville, Särndal et Sautory

Plusieurs algorithmes produisent des coefficients de pondération positifs avec une probabilité élevée. Les méthodes itératives du quotient (raking ratio) produisent des coefficients de pondération positifs pour la plupart des configurations de données. Deville, Särndal et Sautory

où $f(\mathbf{x}_i; \beta)$ est la valeur estimée du modèle pour la $i^{\text{ème}}$ observation.

Firth et Bennett (1998) ont fait remarquer que certains modèles non linéaires satisfont (7.1). Si le modèle initial ne satisfait pas (7.1), nous pouvons ajouter un terme estimatif de coordonnée à l'origine afin de créer un modèle complet étendu.

(6.6)

$$\sum_{i \in A} [w_i - \alpha_i - \alpha_i \log \alpha_i^{-1} w_i].$$

et pour obtenir les coefficients de pondération par le maximum de vraisemblance, ils ont utilisé la fonction objective

Deville, Särndal et Sautory (1993) ont étudié quatre estimateurs dans la classe. Bien que les coefficients de pondération construits en se servant de diverses fonctions puissent différer considérablement, ces auteurs ont conclu que les estimations sont fort similaires, résultat qui corrobore la théorie. Singh et Mohl (1996) et Thøgers (1999, 2000) discutent des estimateurs ayant la propriété de calage.

7. POPULATION DE VECTEURS AUXILIAIRES CONNUS À L'ÉTAPE DE L'ESTIMATION

Si l'on connaît le vecteur \mathbf{x} pour tous les éléments de la

population, le nombre d'estimateurs par régression possibles augmente considérablement. La plupart des méthodes comprenant l'ajustement d'une fonction auxiliaire représentent la relation entre y et les variables auxiliaires. La méthode la plus courante consiste à affecter les éléments de la population à des catégories d'après les données auxiliaires et à utiliser ces catégories comme strates à posteriori. Cette méthode équivaut à approximer l'espérance de y , étant donné \mathbf{x} , au moyen d'une fonction en escalier. L'estimateur est formellement équivalent à l'estimateur par régression (4.19) où le vecteur \mathbf{x} est un vecteur de variables indicatrices de l'appartenance à la strate à posteriori.

L'application de la méthode nécessite souvent l'établissement des critères à utiliser pour former les strates à posteriori. Habituellement, celles-ci sont formées de sorte que chacune contienne un nombre minimal d'éléments de l'échantillon et que les coefficients de pondération pour toute strate à posteriori ne soit pas indûment grand. L'estimation par stratification a posteriori et la formation de strates à posteriori ont été étudiées, entre autres, par Fuller (1966), Holt et Smith (1979), Tremblay (1986), Kalton et Magillig (1991), Little (1993), Eltinge et Yansaneh (1997) et Lazzeroni et Little (1998). Holt et Smith (1979) ont présenté des arguments en faveur de l'utilisation d'un estimateur conditionnel de la variance pour la stratification a posteriori.

Étant donné les vecteurs \mathbf{x} de la population, nous pouvons utiliser l'échantillon pour estimer une relation fonctionnelle entre y et \mathbf{x} , puis prédire la valeur non observée de y . Si l'on veut que la méthode soit convergente par rapport au plan d'échantillonnage, il faut qu'une condition comparable à (4.14) soit satisfaite. Un moyen d'assurer la convergence par rapport au plan d'échantillonnage est de contraindre le modèle ajusté à satisfaire.

$$\sum_{i \in A} \pi_i^{-1} [y_i - f(\mathbf{x}_i; \beta)] = 0, \quad (7.1)$$

Il s'agit d'une extension directe de la notion d'estimation par différence au cas non linéaire. Consulter Tsaki (1970), Cassel, Särndal et Wretman (1976) et Wright (1983). Wu et Sitter (2001) ont proposé une méthode étroitement liée qui consiste à utiliser la fonction ajustée $f(\mathbf{x}_i; \beta)$ comme variable auxiliaire dans un estimateur par régression linéaire.

Outre les fonctions en escalier, plusieurs méthodes « locales » peuvent être utilisées pour approximer la relation fonctionnelle entre \mathbf{x} et y . Les fonctions splines et les polynômes sont des modèles linéaires qui rentrent dans la classe de la section 4. Kuo (1988), Dorfman (1993), Dorfman et Hall (1993), Chambers (1996) et Chambers, Dorfman et Wehrly (1993) ont considéré, pour les populations finies, du point de vue du modèle, des estimateurs comprenant une forme ou l'autre de lissage local pour estimer les quantités de population. Bredt et Opsomer (2000) ont montré que les estimateurs fondés sur une régression polynomiale locale sont convergents par rapport au plan d'échantillonnage. Firth et Bennett (1998) ont, eux aussi, considéré des modèles à ajustement local.

8. ESTIMATION PAR RÉGRESSION ET NON-RÉPONSE

L'estimation par régression fait souvent partie des méthodes utilisées pour corriger les données pour la non-réponse partielle. On peut se fonder, pour justifier la régression, sur un modèle tel que (3.1) ou sur le fait que la régression permet de faire la correction pour tenir compte des probabilités inégales de réponse. Voir Cassel, Särndal et Wretman (1979, 1983), Little (1982, 1986), Bethlehem (1988), Koit (1994), Fuller, Louglin et Baker (1994) et Fuller et An (1998).

Considérons un estimateur du vecteur de coefficients de régression de la population de la forme (4.4) quand $\Phi = \mathbf{D}$ construit en se fondant sur les unités répondantes. Représentons l'estimateur par β et supposons que p_i est la probabilité conditionnelle d'observer l'unité i , étant donné que l'unité est sélectionnée dans l'échantillon. Alors, dans les conditions de régularité, l'estimateur β est un estimateur convergent de

pour cela, utiliser un estimateur de la variance basé sur un modèle,

$$V\{\hat{\beta}^{*h} | \mathbf{H}\} = \left(\mathbf{H}' \hat{\Sigma}^{-1} \mathbf{H} \right)^{-1}$$

ou l'estimateur basé sur le plan d'échantillonnage de la variance de (4.12). Consulter Du Mouchel et Duncan

(1983) et Fuller (1984).

Pour les échantillons à deux degrés, une spécification de

travail pour Σ_{ee} pourrait être particulièrement indiquée; à

cet égard, consulter Royall (1976, 1986) et

Montanari (1987). Un modèle raisonnable est un modèle

dans lequel il existe une corrélation commune entre les

éléments d'une même unité primaire d'échantillonnage,

mais une corrélation nulle entre les éléments contenus dans

des unités primaires d'échantillonnage différentes. Comme

la matrice des covariances associée Σ_{ee} est une matrice dia-

gonale par blocs d'une forme particulière, il est assez facile

de l'inverser et, donc, l'estimateur fondé sur une telle spéci-

fication Φ de travail est assez facile à construire. L'estima-

teur par régression obtenu lorsque l'on utilise une spéci-

fication Φ telle que la corrélation n'est pas nulle pour les

éléments compris dans une même unité primaire d'échan-

tillonnage est une combinaison de l'estimateur fondé sur les

totaux des unités primaires d'échantillonnage et celui fondé

sur les éléments. Voir Fuller et Battese (1973). Donc,

l'utilisation d'une telle spécification Φ permet d'éviter les

problèmes de variance associés à l'utilisation des totaux

correspondant aux unités primaires d'échantillonnage.

6. MAXIMUM DE VRAISEMBLANCE ET MÉTHODE ITÉRATIVE DU QUOTIENT (RAKING RATIO)

Le fondement théorique des estimateurs par régression décrits aux sections 3 et 4 est l'estimation du maximum de vraisemblance dans le cas du modèle linéaire dont les erreurs obéissent à la loi normale. Nous considérons maintenant le maximum de vraisemblance dans le cas de variables multinomiales. Étant donné un échantillon aléatoire simple provenant d'une multinomiale définie par les entrées dans un tableau à double entrée, le logarithme du rapport des vraisemblances est, à une constante près,

$$(6.1) \quad \sum_{i=1}^I \sum_{j=1}^J a_{ij} \log p_{ij},$$

où a_{ij} est la fraction estimée dans la cellule ij , p_{ij} est la fraction de la population dans la cellule ij , r est le nombre de lignes et c est le nombre de colonnes. Si l'on maximise (6.1) en imposant la contrainte $\sum_{j=1}^J p_{ij} = 1$, on obtient les estimateurs du maximum de vraisemblance $\hat{p}_{ij} = a_{ij}^*$. Si l'on connaît les fractions de marge des lignes $p_{i \cdot}^*$ et les fractions de marge des colonnes $p_{\cdot j}^*$, il est naturel de maximiser la vraisemblance subordonnée à ces contraintes en utilisant le lagrangien

$$\hat{\mathbf{p}} = \left[\sum_{i \in A} \mathbf{x}_i' \Phi_{ii}^{-1} \mathbf{x}_i \right]^{-1} \sum_{i \in A} \mathbf{x}_i' \Phi_{ii}^{-1} \mathbf{y}_i,$$

où

$$(6.4) \quad \mathbf{y}^{reg} = \mathbf{y}^* + (\mathbf{x}^* \mathbf{x}^* - \mathbf{x}^* \mathbf{y}^*),$$

moyenne de la forme

est donné par un estimateur par régression de la

une approximation de la solution du problème de mini-

satisfont (6.3). Si le coefficient de pondération initial est

$\alpha_i = (\sum_{j=1}^J \pi_{ij}^{-1})^{-1} \pi_{ij}^{-1}$ et que \mathbf{I} est le premier élément de \mathbf{x}_i ,

étalonné pour décrire les coefficients de pondération qui

Deville et Särndal (1992) ont utilisé le terme *calé* ou

$$(6.3) \quad \sum_{i \in A} w_i' \mathbf{x}_i = \mathbf{x}^* N.$$

la contrainte

final w_i . La fonction objective est minimisée en imposant

de pondération initial α_i et un coefficient de pondération

$G(w, \alpha)$ est une mesure de la distance entre le coefficient

fonctions objectives de la forme $\sum_{i \in A} G(w_i, \alpha_i)$, où

Deville et Särndal (1992) ont considéré une classe de

Bishop, Fienberg et Holland (1975, ch. 3).

à Deming et Stephan (1940). Consulter, par exemple,

suite. La paternité de la méthode est généralement attribuée

ratios tenant compte des contraintes de lignes, et ainsi de

compte des contraintes de colonnes, puis des ajustements de

contraintes de lignes, puis des ajustements de ratios tenant

faire des ajustements de ratios tenant compte des

Il s'agit d'un processus d'itération où l'on commence par

quotient (raking ratio) ou ajustement proportionnel itératif.

vraisemblance est celle appelée *méthode itérative du*

matons s'approchant de la solution du maximum de esti-

vides est trop élevée. Une méthode qui produit des esti-

pourrait qu'il n'y en ait aucune si le nombre de cellules

aucune expression explicite de la solution de (6.2) et il se

et γ_j , $j = 1, 2, \dots, c$, des contraintes de colonnes. Il n'existe

où λ_i , $i = 1, 2, \dots, r$, tient compte des contraintes de lignes

$$(6.2) \quad \left(\sum_{j=1}^J \lambda_j \left(\sum_{i=1}^{r+j-1} d_{ij} - p_{j \cdot}^* \right) + \sum_{i=1}^r \lambda_i \left(\sum_{j=1}^J d_{ij} - p_{i \cdot}^* \right) \right)$$

et Φ_{ii} est la dérivée seconde de $G(w, \alpha)$ par rapport à w_i , calculée à $(w_i, \alpha) = (\alpha_i, \alpha_i)$. En suivant cette méthode, Deville et Särndal (1992) ont montré que les estimateurs du maximum de vraisemblance et par la méthode itérative du quotient ont les mêmes distributions limites que l'estimateur par régression (4.18) quand $\Phi = \mathbf{D}^*$. Pour obtenir les coefficients de pondération par la méthode itérative du quotient, ils ont utilisé la fonction objective

$$(6.5) \quad \sum_{i \in A} w_i' \log \alpha_i^{-1} w_i + \alpha_i - w_i \left[\right],$$

L'estimateur (5.3) se réduit à $\mathbf{x}_N^T \mathbf{b}$ s'il existe un η tel

$$\mathbf{X}^A \eta = \Sigma^{eAA} \mathbf{J}^n + \Sigma^{eAA} \mathbf{J}^{N-n}, \quad (5.5)$$

pour tous les échantillons pour lesquels la probabilité est positive. S'il existe aussi une valeur de γ telle que

$$\mathbf{X}^A \gamma = \Sigma^{eAA} \mathbf{D}^{-1} \mathbf{J}^n \quad (5.6)$$

pour tous les échantillons pour lesquels la probabilité est positive, alors l'estimateur θ de (5.3) est convergent par rapport au plan d'échantillonnage, où \mathbf{D}^n est défini pour

(4.14). Étant donné un \mathbf{k} tel que

$$\mathbf{X}^A \mathbf{k} = \Sigma^{eAA} (\mathbf{D}^{-1} \mathbf{J}^n - \mathbf{J}^n) - \Sigma^{eAA} \mathbf{J}^{N-n}, \quad (5.7)$$

alors l'estimateur θ de (5.3) peut s'exprimer sous la forme

$$\theta = \bar{y}^n + (\bar{\mathbf{x}}^N - \bar{\mathbf{x}}^n)^T \mathbf{b} \quad (5.8)$$

et, si le plan d'échantillonnage est tel que \bar{y}^N est convergent par rapport au plan d'échantillonnage pour \bar{y}^N , l'estimateur θ de (5.8) est convergent par rapport au plan d'échantillonnage pour \bar{y}^N .

Nous appelons un modèle de régression de la forme (5.1) pour lequel (5.5) et (5.6), ou (5.7), comprend un modèle complet. Si la condition (5.6) ou (5.7) n'est pas vérifiée, nous considérons que le modèle est réduit ou restreint. Nous ne pouvons pas nous attendre à ce que les conditions d'un modèle complet soient vérifiées pour chaque variable analysée dans le cas d'une enquête générale, car Σ^{ee} sera

différente pour diverses variables y . Par conséquent, étant donné un modèle réduit, on pourrait rechercher un bon estimateur du modèle dans la classe des estimateurs convergents par rapport au plan d'échantillonnage.

Pour construire un estimateur convergent par rapport au plan d'échantillonnage de la forme $\mathbf{x}_N^T \mathbf{b}$ lorsque le modèle (5.1) est un modèle réduit, nous pouvons ajouter un vecteur satisfaisant (5.7) à la matrice \mathbf{X} pour créer un modèle complet. Dans ce cas, deux situations peuvent se présenter. Dans la première, on connaît la moyenne (ou le total) de la population de la variable ajoutée. Connaissant la moyenne, nous pouvons construire l'estimateur par régression habituel et appliquer les formules habituelles d'estimation de la variance liée au plan d'échantillonnage.

Pour décrire une méthode d'estimation applicable au cas où la moyenne de la population de la variable est inconnue, représentons par $\mathbf{q} = (q_1, q_2, \dots, q_n)^T$ le vecteur ajouté, où \mathbf{q} est le vecteur figurant dans le deuxième membre de l'équation (5.7). Supposons que $\mathbf{H} = (\mathbf{X}, \mathbf{q})$, où \mathbf{X} est la matrice des variables auxiliaires dont on connaît le vecteur des moyennes de la population, $\bar{\mathbf{x}}_N$. Nous écrivons le modèle complet pour l'échantillon sous la forme

$$\mathbf{y} = \mathbf{H} \mathbf{b}_{y \cdot h} + \mathbf{e}, \quad (5.9)$$

où $\mathbf{e} \sim (0, \Sigma^{ee})$. Le meilleur estimateur linéaire conditionnellement sans biais de $\mathbf{b}_{y \cdot h}$ est

$$\hat{\mathbf{b}}_{y \cdot h} = (\mathbf{H}' \Sigma^{ee-1} \mathbf{H})^{-1} \mathbf{H}' \Sigma^{ee-1} \mathbf{y}. \quad (5.10)$$

Dans (5.9), si le coefficient de \mathbf{q} n'est pas nul, il est impossible de construire un estimateur conditionnellement sans biais de $\mathbf{H}^N \mathbf{b}_{y \cdot h}$, car la composante \bar{q}^N de \mathbf{H}^N est inconnue. Cependant, comme l'estimateur $\hat{\mathbf{b}}_{y \cdot h}$ est sans biais, il est possible de construire un estimateur conditionnellement sans biais de tout le terme \bar{q}^N par son « meilleur estimateur disponible » et raisonnable de choisir l'estimateur par régression

$$\bar{q}^{\text{reg}} = \bar{q}^n + (\bar{\mathbf{x}}^N - \bar{\mathbf{x}}^n)^T \hat{\mathbf{b}}_{q \cdot x}, \quad (5.11)$$

où $\hat{\mathbf{b}}_{q \cdot x} = (\mathbf{X}^N \Sigma^{ee-1} \mathbf{X}^N - \mathbf{X}^n \Sigma^{ee-1} \mathbf{X}^n)^{-1} \mathbf{X}^N \Sigma^{ee-1} \mathbf{q}$. Alors, l'estimateur (5.3) devient

$$\theta = \bar{y}^n + (\bar{\mathbf{x}}^N - \bar{\mathbf{x}}^n)^T \hat{\mathbf{b}}_{y \cdot h} \quad (5.12)$$

Nous pouvons exprimer l'estimateur (5.12) sous la forme bien connue de l'estimateur par régression,

$$\bar{y}^{\text{reg}} = \bar{y}^n + (\bar{\mathbf{x}}^N - \bar{\mathbf{x}}^n)^T \hat{\mathbf{b}}_{y \cdot x}. \quad (5.13)$$

Autrement dit, l'estimateur par régression de la moyenne de la population finie pour y fondé sur le modèle complet, lorsque l'on ne connaît pas la moyenne de q_i et qu'on l'estime au moyen de l'estimateur par régression, est l'estimateur par régression où $\mathbf{b}_{y \cdot x}^{\text{reg}}$ est estimé par régression de y sur \mathbf{x} par les moindres carrés généralisés en utilisant la matrice des covariances Σ^{ee} . Consulter Park (2002). L'estimateur est conditionnellement sans biais par rapport au modèle dans le cas du modèle réduit contenant uniquement \mathbf{x} si le modèle réduit est valable. Si le coefficient de population pour q_i n'est pas nul, le modèle réduit n'est pas correct. Alors, l'estimateur est conditionnellement biaisé par rapport au modèle, mais il est sans biais pour la moyenne de la population finie dans les conditions du modèle complet et d'un plan d'échantillonnage sans biais.

$$E\{\bar{y}^{\text{reg}} - \bar{y}^N | \mathbf{H}\} = E\{E[\bar{y}^{\text{reg}} - \bar{y}^N | \mathbf{H}]\} = 0, \quad (5.14)$$

où \bar{y}^{reg} est défini dans (5.12) et l'approximation est due à l'espérance approximative concernant le plan d'échantillonnage dans l'estimateur par régression \bar{q}^{reg} .

L'estimateur (5.13) est un estimateur linéaire, où le vecteur des coefficients de pondération, \mathbf{w} , minimise le lagrangien

$$\mathbf{w}' \Sigma^{ee} \mathbf{w} + [\mathbf{w}' \mathbf{H} - (\bar{\mathbf{x}}^N, \bar{q}^{\text{reg}})] \lambda. \quad (5.15)$$

L'estimateur est invariant en localisation si la colonne de 1 est comprise dans l'espace colonne de \mathbf{X} . Puisque q est la variable dont l'omission dans le modèle complet peut produire un biais, il semble prudent de tester le coefficient de q avant d'utiliser le modèle réduit pour construire un estimateur de y . Nous pouvons,

population. L'estimateur (4.17), tel que $\Phi^{-1} = \text{diag}\{K_1', \dots, K_r'\}$, est sur la diagonale pour les éléments de la strate g , et contenant des variables indicatrices pour les effets de strates, donne l'estimateur de la moyenne dans la classe

$$\bar{y}^{\text{reg}} = \bar{y}_g + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n}) \hat{\beta}_1$$

caractérisé par l'estimation la plus faible de la variance liée au plan d'échantillonnage. Si les pentes réelles dans les strates sont identiques et que les probabilités de sélection sont proportionnelles à la racine carrée des variances à l'intérieur des strates, alors l'utilisation de $\Phi = D_2^{\pi^*}$ produit l'intérieur des strates, une EQM plus faible que l'utilisation de $\Phi^{-1} = \text{diag}\{K_1', \dots, K_r'\}$, car la somme des $w_{hi}^2 \sigma_h^2$ est plus faible. Fuller et Isaki (1981) ont remarqué que l'estimateur optimal par rapport au plan d'échantillonnage est souvent bien approximé par l'estimateur construit en prenant $\Phi = D_2^{\pi^*}$.

Nous avons introduit une estimation par régression de la moyenne, mais ce sont souvent les totaux que l'on estime et que l'on utilise comme contrôles. Considérons l'estimateur par régression du total de y défini par

$$\hat{T}_{y,\pi}^{\text{reg}} = \hat{T}_{y,\pi} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \hat{\beta}_{y \cdot x} \quad (4.33)$$

où $\mathbf{T}_{x,N}$ est le total connu de \mathbf{x} et $(\hat{T}_{y,\pi}, \hat{\mathbf{T}}_{x,\pi})$ est le vecteur des estimateurs de $(T_{y,N}, \mathbf{T}_{x,N})$ convergents par rapport au plan d'échantillonnage. Par analogie à (4.28), l'estimateur

$$\hat{\beta}_{y \cdot x} = [V\{\hat{\mathbf{T}}_{x,\pi}\}]^{-1} C\{\hat{\mathbf{T}}_{x,\pi}, \hat{T}_{y,\pi}\}, \quad (4.34)$$

où $V\{\hat{\mathbf{T}}_{x,\pi}\}$ est un estimateur convergent par rapport au plan d'échantillonnage de la variance de $\hat{\mathbf{T}}_{x,\pi}$ et $C\{\hat{\mathbf{T}}_{x,\pi}, \hat{T}_{y,\pi}\}$ est un estimateur convergent par rapport au plan d'échantillonnage de la covariance de $\hat{\mathbf{T}}_{x,\pi}$ et $\hat{T}_{y,\pi}$.

L'estimateur du total est $N \bar{y}^{\text{reg}}$ dans le cas de l'échantillonnage aléatoire simple, mais l'équivalence exacte pourrait ne pas être vérifiée dans le cas d'échantillons plus complexes, car l'estimateur de la moyenne pourrait être un estimateur par ratio. Cependant, si nous construisons l'estimateur par régression des deux totaux au moyen de (4.34), la variance pour grands échantillons du ratio des deux totaux est égale à l'estimateur par régression de la

$$\hat{T}_{y,\pi}^{\text{reg}} - T_{y,N} = \hat{T}_{y,\pi} - T_{y,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \hat{\beta}_{y \cdot x} + O_p(N^{-1})$$

et

$$\hat{T}_{u,\text{reg}} - T_{u,N} = \hat{T}_{u,\pi} - T_{u,N} + (\mathbf{T}_{x,N} - \hat{\mathbf{T}}_{x,\pi}) \hat{\beta}_{u \cdot x} + O_p(N^{-1}), \quad (4.35)$$

où $V(\hat{\mathbf{e}}_i^{\pi})$ est l'estimateur de (4.26) construit en prenant dans le contexte des grands échantillons, (4.29) montre

$$C\{\bar{\mathbf{x}}_{1,\pi}, \pi\} = \sum_{h=1}^H K_h^{\pi} \sum_{n_h}^f (\mathbf{x}_{1,h} - \bar{\mathbf{x}}_{1,h})' (\mathbf{y}_h - \bar{y}_h), \quad (4.31)$$

$$K_h = W_h^2 (1 - f_h) (n_h - 1)^{-1} n_h^{-1} = N^{-2} \pi_h^2 (1 - f_h) (n_h - 1)^{-1} n_h^{-1}$$

$N^{-1} N_h = W_h, N_h$ est la taille de la strate h , $f_h = \pi_h = N^{-1} n_h$ et n_h est la taille de l'échantillon dans la strate h . Il s'ensuit que les coefficients de pondération associés à l'estimateur (4.29) sont

$$W_{hi} = N^{-1} \pi_{hi}^{-1} + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})' \times \left[\sum_{j=1}^H K_j' \sum_{n_j}^f (\mathbf{x}_{1,j} - \bar{\mathbf{x}}_{1,j})' (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,i}) \right]^{-1} \times K_h (\mathbf{x}_{1,hi} - \bar{\mathbf{x}}_{1,h})', \quad (4.32)$$

Voir aussi Särndal (1996). Nous pouvons construire les coefficients de pondération de (4.32), à condition que soient satisfaites les contraintes

$$\sum_{h=1}^H W_{hi} = N^{-1} n_h, \quad h = 1, 2, \dots, H,$$

où A_h représente l'ensemble d'éléments dans la strate h . L'estimateur de (4.19) quand $\Phi = D_2^{\pi^*}$ est une fonction des estimateurs d'Horvitz-Thompson des moments de la

Comme l'estimateur par régression de la moyenne est une combinaison linéaire de coefficients de régression, il s'agit du coefficient de régression d'une combinaison linéaire des variables x originales. Pour montrer ceci, supposons $\mathbf{x}_i = (x_{0,i}, \mathbf{x}_{1,i})' = (1, \mathbf{x}_{1,i})'$, et définissons un nouveau vecteur dont la première valeur est 1 et un deuxième vecteur dont la moyenne de la population est nulle obtenue en soustrayant la moyenne de population originale $\bar{\mathbf{x}}_{1,N}$ du vecteur original $\mathbf{x}_{1,i}$. Représentons par $\mathbf{q}_i = (1, \mathbf{x}_{1,i})'$ le vecteur transformé. Alors, le modèle de régression

$$y_i = \mathbf{q}_i' \boldsymbol{\gamma} + e_i, \quad (4.21)$$

où le vecteur de coefficients pour la population finie est

$$\boldsymbol{\gamma}_N = (\bar{y}_N, \boldsymbol{\beta}_{1,N}')' = \left(\sum_{i \in U} \mathbf{q}_i' \mathbf{q}_i \right)^{-1} \sum_{i \in U} \mathbf{q}_i' y_i. \quad (4.22)$$

L'expression de l'estimateur par régression de la moyenne devient

$$\bar{y}_{\text{reg}} = \bar{\mathbf{q}}_N' \boldsymbol{\gamma} = \bar{y}_0, \quad (4.23)$$

où \bar{y} est obtenu d'après (4.4) en remplaçant \mathbf{x}_i par \mathbf{q}_i . Puisqu'il s'agit d'un estimateur linéaire de la forme $\mathbf{w}'\mathbf{y}$, nous pouvons écrire

$$\bar{y}_{\text{reg}} = \sum_{i \in A} w_i y_i = \sum_{i \in A} \pi_i^{-1} g_i y_i, \quad (4.24)$$

où $w_i = \pi_i^{-1} g_i$. De surcroît, la variance estimée d'après (4.12) est

$$V\{\bar{y}_{\text{reg}}\} = V\left\{\bar{y}_0\right\} = V\left\{\sum_{i \in A} \pi_i^{-1} (g_i e_i)\right\}, \quad (4.25)$$

où il est sous-entendu que l'estimation de la variance liée au plan d'échantillonnage de (4.25) est calculée pour la variable $g_i e_i$, $e_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$, et que $\boldsymbol{\beta}$ est défini dans (4.4). L'estimateur de la variance (4.25) est une généralisation directe de l'expression (3.5). Si nous transformons la variable de sorte que la moyenne de population du vecteur de auxiliaire soit nulle, le premier élément du vecteur de coefficients de régression est la régression de la moyenne et le premier élément de (4.12) est un estimateur de la variance de l'estimateur par régression qui contient une composante due à l'estimation de $\boldsymbol{\beta}$. Ce fait a été mentionné dans Hidiroglou, Fuller et Hickman (1978). Consulter aussi Särndal (1982), Särndal, Swensson et Wretman (1989) ont proposé la terminologie du facteur g pour le calcul de la variance estimative d'un total estimé par régression.

Partant de (4.17), nous pouvons écrire

$$\bar{y}_{\text{reg}} = \bar{\mathbf{x}}_{0,N}' \bar{\mathbf{x}}_{0,1}^{-1} \bar{\mathbf{y}} - (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \boldsymbol{\beta}_{1,N})' (\bar{\mathbf{y}}_N - \bar{\mathbf{x}}_{1,N} \boldsymbol{\beta}_{1,N})^{-1} + O_p(n^{-1}),$$

où $e_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$. Donc, nous pouvons estimer la variance de l'estimateur par régression au moyen de

$$V\{\bar{e}_n\} = V\left\{\sum_{i \in A} \pi_i^{-1} \bar{e}_i\right\} \quad (4.26)$$

où $\bar{e}_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$. Nous recommandons d'utiliser l'estimateur (4.25), puisqu'il est aussi facile à calculer que l'estimateur (4.26) et qu'il est applicable lorsque $\bar{\mathbf{x}}_{1,n} - \bar{\mathbf{x}}_{1,N}$ n'est pas $O_p(n^{-1/2})$.

Nous pouvons aussi calculer la variance de l'estimateur par régression par la méthode du jackknife ou par d'autres méthodes de répétition qui sont utilisées de plus en plus fréquemment. Consulter Frankel (1971), Kish et Frankel (1974), Woodruff et Causey (1976), Royall et Cumberland (1978), et Duchesne (2000). Yung et Rao (1996) ont montré que l'estimateur (4.25) est identique à un estimateur par linéarisation jackknife dans le cas des plans d'échantillonnage stratifiés à plusieurs degrés.

La méthode d'estimation par régression associée à (4.18) et (4.19) s'inscrit entièrement dans le cadre de la formulation d'un plan d'échantillonnage. Outre l'existence des moments, aucun modèle de population n'est utilisé, bien qu'il puisse être soutenu que le recours à la régression n'est sans doute envisagé que s'il existe une certaine corrélation linéaire entre $\mathbf{x}_{1,i}$ et y_i . L'estimateur (4.19) est un estimateur fort naturel, car l'estimateur du coefficient de régression est un estimateur convergent par rapport au plan d'échantillonnage du coefficient de régression de la population. Il est un peu frustrant que (4.18) ne produise pas systématiquement la valeur la plus faible de la variance liée au plan d'échantillonnage pour les grands échantillons pour l'estimation de la moyenne. Si nous traitons $\boldsymbol{\beta}_1$ dans (4.18) comme un vecteur fixe, la valeur qui minimise la variance de la combinaison linéaire des moyennes est

$$\boldsymbol{\beta}_{1,\text{dopt}} = [V\{\bar{\mathbf{x}}_{1,n} | \mathbf{F}_N\}^{-1} C\{\bar{\mathbf{x}}_{1,n}, \bar{\mathbf{y}}_n | \mathbf{F}_N\}]^{-1}. \quad (4.27)$$

À cet égard, consulter Cochran (1977, page 201), Fuller et Isaki (1981), Montanari (1987, 1999) et Rao (1994). S'il existe un estimateur de la variance de $\bar{\mathbf{x}}_{1,n}$ convergent par rapport au plan d'échantillonnage, alors le $\boldsymbol{\beta}_{1,d}$ qui minimise l'estimateur de la variance

$$V\{\bar{\mathbf{y}}_n - \bar{\mathbf{x}}_{1,n} \boldsymbol{\beta}_{1,d}\}, \quad (4.28)$$

représenté par $\boldsymbol{\beta}_{1,\text{dopt}}$ est un estimateur convergent de $\boldsymbol{\beta}_{1,\text{dopt}}$. Il s'ensuit que l'estimateur

$$\bar{y}_{\text{d,reg}} = \bar{\mathbf{y}}_n + (\bar{\mathbf{x}}_{1,n} - \bar{\mathbf{x}}_{1,N})' \boldsymbol{\beta}_{1,\text{dopt}} \quad (4.29)$$

est caractérisé par la variance limite minimale pour les estimateurs convergents par rapport au plan d'échantillonnage de la forme $\bar{\mathbf{y}}_n + (\bar{\mathbf{x}}_{1,n} - \bar{\mathbf{x}}_{1,N})' \boldsymbol{\beta}_{1,d}$. En outre

$$V\{\bar{e}_n\} = \bar{\mathbf{y}}_n' (\bar{\mathbf{y}}_n - \bar{\mathbf{y}}_N) \bar{\mathbf{y}}_n^{-1} \bar{\mathbf{y}}_n' N(0, 1), \quad (4.30)$$

rapport au plan d'échantillonnage de \bar{y}_N . Il découle de (4.11) que

$$\left[\bar{x}_N^N V \{ \bar{b} \} \bar{x}_N^N \right]^{-1/2} \left(\bar{x}_N^N \bar{b} - \bar{y}_N \right) \xrightarrow{L} N(0, I). \quad (4.15)$$

L'exigence (4.14) selon laquelle $\Phi D_{-1}^N \mathbf{f}$ doit se trouver dans l'espace colonne de \mathbf{X} , est une condition essentielle à la convergence par rapport au plan d'échantillonnage. Un moyen simple de satisfaire cette exigence consiste à permettre qu'une colonne de \mathbf{X} soit une colonne de 1 et à utiliser un multiple de D_{-1}^N pour Φ , ou à permettre qu'une colonne de \mathbf{X} comprenne les éléments π_i^{-1} et fixer $\Phi = \mathbf{I}$, ou à permettre qu'une colonne de \mathbf{X} contienne les éléments π_i et fixer $\Phi = D_2^N$. Si \mathbf{X} est composée du vecteur colonne unique contenant les éléments π_i et que $\Phi = D_2^N$, alors l'estimateur (4.13) se réduit à l'estimateur d'Horvitz-Thompson (4.5) pour des plans d'échantillonnage à taille fixe. Si $\mathbf{X} = \mathbf{J}$ et $\Phi = D_{-1}^N$, l'estimateur (4.13) se réduit à l'estimateur par ratio.

qui est invariant en fonction de la localisation.

Pour déterminer la nature de l'estimateur lorsque la condition (4.14) est satisfaite, supposons, sans perte de généralité, que $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}_1)$, où $\mathbf{x}_0 = \Phi D_{-1}^N \mathbf{f}$ et

$$\mathbf{x}_1 = (\mathbf{x}_{0,i}, \mathbf{x}_{1,i}). \text{ Alors}$$

$$\bar{y}_{reg} = \bar{x}_{0,N} \bar{x}_{0,N}^{-1} \bar{y}_N + (\bar{x}_{1,N} - \bar{x}_{0,N} \bar{x}_{0,N}^{-1} \bar{x}_{1,N}) \bar{b}_1, \quad (4.17)$$

où

$$\bar{b}_1 = \left[(\mathbf{X}_1 - \mathbf{x}_0 \bar{x}_{0,N}^{-1} \bar{x}_{1,N})' \Phi^{-1} (\mathbf{X}_1 - \mathbf{x}_0 \bar{x}_{0,N}^{-1} \bar{x}_{1,N}) \right]^{-1} \times (\mathbf{X}_1 - \mathbf{x}_0 \bar{x}_{0,N}^{-1} \bar{x}_{1,N})' \Phi^{-1} \bar{y}_N,$$

$\bar{b}_{x_1} = \bar{x}_{-1}^{-1} \bar{x}_{-1} \bar{x}_{-1}^{-1} \bar{x}_{-1}$, et $(\bar{y}_N, \bar{x}_{-1})$ est défini dans (4.16). Les ratios, tels que $\bar{x}_{0,N}^{-1} \bar{y}_N$, peuvent également s'écrire sous forme de ratios d'estimateurs d'Horvitz-Thompson. Si \mathbf{J} se trouve dans l'espace colonne de \mathbf{X} , l'estimateur (4.17) est invariant en fonction de la localisation. Si $\Phi = D_{-1}^N$, alors

$$\bar{y}_{reg} = \bar{x}_N^N \bar{b} + (\bar{x}_{1,N} - \bar{x}_N^N \bar{x}_{1,N}^{-1} \bar{x}_N^N) \bar{b}_1, \quad (4.18)$$

où

$$\bar{b}_1 = \left[\sum_{i \in U} (\mathbf{x}_{1,i} - \bar{x}_{1,N} \pi_i^{-1} (\mathbf{x}_{1,i} - \bar{x}_{1,N} \pi_i^{-1})) \right]^{-1} \times \sum_{i \in U} (\mathbf{x}_{1,i} - \bar{x}_{1,N} \pi_i^{-1} (\mathbf{x}_{1,i} - \bar{x}_{1,N} \pi_i^{-1}))' \pi_i^{-1} (\mathbf{y}_i - \bar{y}_N).$$

En outre, quand $\Phi = D_{-1}^N$, le \bar{b}_N de (4.7) est le coefficient de régression pour la population

$$\bar{b}_N = \left[\sum_{i \in U} \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \sum_{i \in U} \mathbf{x}_i' \mathbf{y}_i. \quad (4.20)$$

Thompson de la moyenne. Dans l'expression (4.6), le \bar{b}_N est défini comme étant une fonction des valeurs attendues des quantités d'échantillons $(\bar{Q}_{xx}, \bar{Q}_{xy})$. Par conséquent, \bar{b}_N n'est pas nécessairement le coefficient de régression par les moindres carrés ordinaires pour la population finie. Le vecteur \bar{b}_j de (4.9) est la généralisation du vecteur \bar{b}_j de (3.3). Si la distribution limite de l'estimateur d'Horvitz-Thompson normalisé correctement est la distribution normale et qu'il existe un estimateur de la variance de cet estimateur, convergent par rapport au plan d'échantillonnage, alors il est possible de construire des tests et des intervalles de confiance pour les coefficients. Supposons que le plan d'échantillonnage soit tel que

$$\bar{V}^{-1/2} (\bar{z}^{HT} - \bar{z}_N) | \mathbf{F}_N \xrightarrow{L} N(0, \mathbf{I}), \quad (4.10)$$

lorsque $N, n \rightarrow \infty$, où \bar{V}^{zz} est la matrice des covariances de $\bar{z}^{HT} - \bar{z}_N$. Si \bar{V}^{zz} est $O(n^{-1})$ et que l'estimateur \bar{V}^{zz} est approprié pour \bar{V}^{zz} , alors

$$\left[\bar{V} \{ \bar{b} \} \right]^{-1/2} (\bar{b} - \bar{b}_N) | \mathbf{F}_N \xrightarrow{L} N(0, \mathbf{I}), \quad (4.11)$$

où

$$\bar{V} \{ \bar{b} \} = \bar{Q}_{-1}^{-1} \bar{V}^{bb} \bar{Q}_{-1}^{-1} = \bar{V} \{ \bar{c}^{HT} \}, \quad (4.12)$$

$\bar{V}^b = \bar{V} \{ \bar{b}^{HT} \}$ est l'estimation de la variance de \bar{b}^{HT} due au plan d'échantillonnage calculée avec $\bar{b}_i' = n^{-1} N \pi_i^{-1} \bar{c}_i'$, $\bar{c}_i' = \mathbf{y}_i - \mathbf{x}_i' \bar{b}_i$, et $\bar{V} \{ \bar{c}^{HT} \}$ est l'estimation de la variance de \bar{c}^{HT} due au plan d'échantillonnage calculée avec $\bar{c}_i' = \bar{Q}_{xx}^{-1} \bar{b}_i'$. Les propriétés limites tiennent pour des échantillons stratifiés et pour des échantillons stratifiés à deux degrés si l'on impose de légères restrictions sur la séquence des populations.

Par analogie à (3.7), nous obtenons un estimateur de la valeur de la fonction estimée de régression à la régression de la moyenne de la population finie par calcul

$$\bar{y}_{reg} = \bar{x}_N^N \bar{b}, \quad (4.13)$$

où \bar{b} est de la forme (4.4) avec une matrice Φ générale. Nous pouvons écrire l'estimateur sous la forme $\bar{w}' \bar{y}$, où le vecteur des coefficients de pondération peut être construit par minimisation du lagrangien

$$\bar{w}' \Phi \bar{w} + (\bar{w}' \mathbf{X} - \bar{x}_N^N) \bar{\lambda}$$

et $\bar{\lambda}$ est le vecteur des multiplicateurs de Lagrange. Si les vecteurs colonnes \bar{y} sont tels que

$$\mathbf{X} \bar{y} = \Phi D_{-1}^N \mathbf{f} \quad (4.14)$$

pour tous les échantillons possibles, où $D_{-1}^N = \text{diag}(\pi_1^{-1}, \pi_2^{-1}, \dots, \pi_n^{-1})$ et \mathbf{f} est un vecteur colonne de 1 à n dimensions, l'estimateur par régression $\bar{x}_N^N \bar{b}$ de (4.13), \bar{b} étant défini dans (4.4), est un estimateur convergent par

4. ESTIMATION FONDÉE SUR LE PLAN D'ÉCHANTILLONNAGE

À la présente section, nous traitons la population finie

population infinie. Dans le domaine de l'échantillonnage, l'utilisation de ce genre de modèle remonte assez loin. Une série de références allant jusqu'à 1970 inclut Cochran (1939, 1942, 1946), Deming et Stephan (1941), Madow et Madow (1944), Yates (1949), Godambe (1955), Hájek (1959), Rao, Hartley et Cochran (1962), Kohnen (1962), Brewer (1963), Godambe et Joshi (1965), Hanurav (1966), Ericson (1969), Isaki (1970) et Royall (1970).

Pour discuter des propriétés des estimateurs par régression dans le cas de grands échantillons, nous considérons des séries de populations finies et d'échantillons probabilistes connexes. L'ensemble d'indices des éléments de la $N^{\text{ème}}$ population finie est $U^N = \{1, \dots, N\}$, où $N = 1, 2, \dots$. Un vecteur ligne de caractéristiques $\mathbf{z}^N = (z_1^N, \dots, z_N^N)$ est associé au $i^{\text{ème}}$ élément de la $N^{\text{ème}}$ population. Supposons que

$$\mathbf{F}^N = [(y_1^N, \mathbf{x}_1^N), (y_2^N, \mathbf{x}_2^N), \dots, (y_N^N, \mathbf{x}_N^N)]$$

est l'ensemble de vecteurs pour la $N^{\text{ème}}$ population finie. L'indice N est souvent omis sur les vecteurs. La moyenne

$$\bar{\mathbf{z}}^N = (\bar{y}^N, \bar{\mathbf{x}}^N) = N^{-1} \sum_{i=1}^N (y_i, \mathbf{x}_i). \quad (4.1)$$

Nous représentons par A^N l'ensemble d'indices qui apparaissent dans l'échantillon sélectionné à partir de la $N^{\text{ème}}$ population finie.

Si la population finie est un échantillon tiré d'une superpopulation infinie, les propriétés probabilistes de l'échantillon sont déterminées d'après les propriétés de la superpopulation et celles de la méthode probabiliste utilisée pour sélectionner l'échantillon. Nous pouvons considérer les propriétés inconditionnelles, les propriétés conditionnelles associées à la population finie particulière ou les propriétés conditionnelles associées à une certaine partie de l'échantillon réalisé.

Les propriétés conditionnelles associées à la population finie dépendent principalement du plan d'échantillonnage et sont souvent appelées propriétés liées au plan d'échantillonnage. Donc, un estimateur $\hat{\theta}$ est dit convergent par rapport au plan d'échantillonnage pour le paramètre de population finie θ_N si, pour tout $\varepsilon > 0$,

$$\lim_{N \rightarrow \infty} \text{prob} \left\{ |\hat{\theta} - \theta_N| > \varepsilon \mid \mathbf{F}^N \right\} = 0,$$

où la notation signifie que nous imposons les contraintes de la population finie réalisée \mathbf{F}^N et, donc, que la probabilité est déterminée d'après le plan d'échantillonnage. Supposons que la population finie est générée par sélections indépendantes à partir d'une superpopulation

pour laquelle $E\{\mathbf{z}_i/\mathbf{z}_i'\}$ est définie et positive, où $\mathbf{z}_i = (y_i, \mathbf{x}_i')$. Nous définissons un vecteur de superpopulation des coefficients de régression par les moindres carrés par

$$\boldsymbol{\beta} = [E\{\mathbf{x}_i' \mathbf{x}_i'\}^{-1} E\{\mathbf{x}_i' y_i'\}] \quad (4.2)$$

Étant donné un échantillon de n observations sur \mathbf{z}_i , nous représentons par $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$, où la $i^{\text{ème}}$ ligne de \mathbf{Z} est (y_i, \mathbf{x}_i') . Si nous supposons le modèle

$$y = \mathbf{X}\boldsymbol{\beta} + u, \quad \mathbf{E}\{u, uu'\} = (\mathbf{0}, \boldsymbol{\Phi}), \quad (4.3)$$

l'estimateur par les moindres carrés généralisés de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}. \quad (4.4)$$

Le modèle (4.3) sert de motif pour les estimateurs de la forme (4.4), mais nous considérerons les estimateurs où $\boldsymbol{\Phi}$ est une matrice générale symétrique définie positive de coefficients de pondération qui n'est pas nécessairement la matrice des covariances des erreurs.

Nous nous inspirons de Fuller (1975) pour donner les propriétés liées aux grands échantillons du vecteur de coefficients estimés de régression (4.4). Consulter aussi Hidiroglou (1974), Scott et Wu (1981), et Robinson et Sandral (1983).

Supposons que la superpopulation possède huit moments et que le plan d'échantillonnage est tel que l'erreur dans l'estimateur d'Horvitz-Thompson de la moyenne est

$$\mathbf{z}^{\text{HT}} = (\mathbf{y}^{\text{HT}}, \mathbf{x}^{\text{HT}}) = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{z}_i \quad (4.5)$$

et π_i est la probabilité de sélection pour l'élément i . Alors, l'erreur dans le vecteur des coefficients de régression est

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \mathbf{F}^N = \mathbf{Q}^{-1} \mathbf{b}^{\text{HT}} + O_p(n^{-1}), \quad (4.6)$$

où

$$\boldsymbol{\beta}^N = \mathbf{Q}^{-1} \mathbf{Q}^{\text{xxN}} \quad (4.7)$$

$$(\mathbf{Q}^{\text{xxN}}, \mathbf{Q}^{\text{xyN}}) = E\left\{(\hat{\mathbf{Q}}^{\text{xx}}, \hat{\mathbf{Q}}^{\text{xy}}) \mid \mathbf{F}^N\right\}, \quad (4.8)$$

$$(\hat{\mathbf{Q}}^{\text{xx}}, \hat{\mathbf{Q}}^{\text{xy}}) = n^{-1} (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}, \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}),$$

$$\mathbf{b}^{\text{HT}} = N^{-1} \sum_{i \in A} \pi_i^{-1} \mathbf{b}_i \quad (4.9)$$

$\mathbf{b}_i' = n^{-1} N \pi_i \zeta_i e_i'$, $e_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}^N$, et ζ_i est la colonne i de $\mathbf{X}'\boldsymbol{\Phi}^{-1}$. L'erreur dans l'estimateur de $\boldsymbol{\beta}^N$ est approximativement égale à l'erreur dans un estimateur d'Horvitz-

où les coefficients de pondération, w_{ai} , minimisent le lagrangien

$$\sum_{i \in A} w_{ai} + \sum_{j=1}^J \lambda_j \left(\sum_{i \in A} w_{ai} x_{ij} - a_j \right)$$

et les λ_j sont des multiplicateurs de Lagrange. La variance de $\hat{\theta}_a$ est

$$V\{\hat{\theta}_a\} = V\left\{ \sum_{i \in A} w_{ai} e_i \right\} = \sum_{i \in A} w_{ai}^2 \sigma_e^2$$

car les coefficients de pondération sont des fonctions de \mathbf{x}_i

et non de y_i .

La matrice des covariances de $\hat{\mathbf{b}}$ est

$$V\{\hat{\mathbf{b}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} V\left\{ \sum_{i \in A} \mathbf{b}_i' \right\} \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$$

$$= V\left\{ \sum_{i \in A} \mathbf{c}_i' \right\} \quad (3.3)$$

où $\mathbf{b}_i' = \mathbf{x}_i' e_i$ et $\mathbf{c}_i' = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i' e_i$. Comme e_i est indépendant de \mathbf{x}_i pour toutes les valeurs de i et de j ,

$$V\left\{ \sum_{i \in A} \mathbf{b}_i' \right\} = \sum_{i \in A} V\{\mathbf{b}_i'\} = \sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \sigma_e^2$$

et nous obtenons l'expression bien connue

$$V\{\hat{\mathbf{b}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sigma_e^2.$$

Nous obtenons l'estimateur sans biais habituel de la matrice des covariances de $\hat{\mathbf{b}}$ en remplaçant σ_e^2 par l'estimateur sans biais de σ_e^2 obtenu par le carré de la moyenne des résidus, $\hat{e}_i = y_i - \mathbf{x}_i' \hat{\mathbf{b}}$. Un estimateur de la matrice des covariances qui estime directement $V\{\sum_{i \in A} \mathbf{b}_i'\}$ est donné par

$$\tilde{V}\{\hat{\mathbf{b}}\} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in A} \mathbf{b}_i' \hat{\mathbf{b}}_i' \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1}$$

$$= \sum_{i \in A} \hat{\mathbf{c}}_i' e_i, \quad (3.4)$$

où $\hat{\mathbf{b}}_i' = \mathbf{x}_i' e_i$ et $\hat{\mathbf{c}}_i' = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i' e_i$. De la même façon

$$V\{\hat{\theta}_a\} = \sum_{i \in A} w_{ai}^2 \hat{e}_i^2 \quad (3.5)$$

est une combinaison linéaire des éléments de (3.4) et un estimateur convergent de $V\{\hat{\theta}_a\}$. L'estimateur (3.4) est un estimateur convergent de $V\{\hat{\mathbf{b}}\}$ si la matrice des covariances des e_i est une matrice diagonale à éléments bornés.

Donc, il s'agit d'un estimateur plus robuste. Cependant, l'estimateur (3.4) est entaché d'un biais par défaut, car la variance de \hat{e}_i est habituellement inférieure à la variance de e_i . Nous disposons de deux méthodes pour réduire le biais. La première consiste à rajuster le nombre de degrés de liberté en multipliant $V\{\hat{\mathbf{b}}\}$ par $(n-k)^{-1}$, où k est la dimension de \mathbf{x}_i . Un autre ajustement consiste à remplacer \hat{e}_i par

$$\hat{e}_i' = (1 - \psi^{(n)})^{-0.5} \hat{e}_i,$$

où $\psi^{(n)}$ représente le terme élément diagonal de $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Consulter Horn, et Duncan (1975), Royall et Cumberland (1978) et Cook et Weisberg (1982, section 2.2).

Si nous observons la valeur \mathbf{x}_i pour un élément, mais que nous n'observons pas y_i , alors le meilleur prédicteur de y_i pour cet élément est $\hat{y}_i = \mathbf{x}_i' \hat{\mathbf{b}}$. Par conséquent, si nous observons la somme des \mathbf{x}_i pour un ensemble de N éléments qui satisfont le modèle (3.1), un sous-ensemble d'observations connues de \mathbf{x}_i pour les $N-n$ éléments restants, l'expression

$$\mathbf{y}_{N-n, \text{reg}} = \sum_{i \in A} \hat{y}_i = \sum_{i \in A} \mathbf{x}_i' \hat{\mathbf{b}},$$

où $\hat{\mathbf{b}}$ représente l'ensemble d'éléments pour lesquels y n'est pas observé, est le meilleur prédicteur de la somme des valeurs non observées de y . Consulter Goldberger (1962), Brewer (1963), Royall (1970), Harville (1976) et Graybill (1976, section 12.2). Par conséquent,

$$\hat{T}_{y, \text{reg}} = \sum_{i \in A} y_i + \mathbf{y}_{N-n, \text{reg}} \quad (3.6)$$

est le meilleur prédicteur du total des N observations. Si le premier élément du vecteur des x est toujours égal à l'unité, nous pouvons procéder à la partition du vecteur des x de sorte que $\mathbf{x}_i = (1, \mathbf{x}_{1,i})'$ et écrire l'estimateur par régression de la moyenne sous la forme

$$\bar{y}_{\text{reg}} = N^{-1} \hat{T}_{y, \text{reg}} = \bar{\mathbf{x}}_N' \hat{\mathbf{b}} = \bar{\mathbf{y}}_n + (\bar{\mathbf{x}}_{1,N} - \bar{\mathbf{x}}_{1,n})' \hat{\mathbf{b}}_1, \quad (3.7)$$

où le $\hat{\mathbf{b}}$ de l'équation (3.2) est partitionné de sorte que $(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1)'$ est le vecteur des simples moyennes d'échantillon. Nous appelons $\bar{\mathbf{x}}_N' \hat{\mathbf{b}}$ l'estimateur par régression des moyennes.

Étant donné le modèle (3.1), l'espérance de la moyenne de y pour la population finie de N éléments gérée par le modèle est $\bar{\mathbf{x}}_N' \hat{\mathbf{b}}$ et $\bar{\mathbf{x}}_N' \hat{\mathbf{b}}$ est un estimateur sans biais de la moyenne de la population finie. Selon nous, cette étape est celle où débute l'estimation par régression de la moyenne d'une population finie dans le cas de plans d'échantillonnage plus complexes.

(2001), Gambino, Kennedy et Singh (2001), et Fuller et Rao (2001).

Bethlehem et Keller (1987) décrivent l'utilisation de l'estimation par régression au Bureau central de la statistique des Pays-Bas (appelé aujourd'hui Statistique Pays-Bas) dans un programme appelé LIN WEIGHT. Nieuwenbroek, Renssen et Hofman (2000) décrivent le projetiel Basculi, qui a remplacé LIN WEIGHT. Deville, Särndal et Sautory (1993) décrivent un programme informatique appelé CALMAR mis au point par l'Institut national de la statistique et des études économiques (INSSEB) qui calcule les coefficients de pondération de type régression en offrant diverses options de pondération objectives. Un programme appelé CLAN97 mis au point par Statistiques Suède est décrit dans Andersson et Nordberg (1998). Enfin, Folsom et Singh (2000) discutent d'une méthode élaborée au Research Triangle Institute.

3. LE MODÈLE LINÉAIRE CLASSIQUE

Le modèle linéaire classique est le fondement de l'estimation par régression à partir de données d'enquête, mais l'application du modèle à ces données nécessite certaines adaptations. En guise d'introduction à l'estimation par régression d'après les données obtenues sur échantillon, nous passons en revue le modèle linéaire classique. Supposons que

$$y_i = x_i'\beta + e_i, \quad i = 1, 2, \dots, n,$$

(3.1)

$$e_i \sim \text{NI}(0, \sigma_e^2),$$

où le terme e_i est indépendant des vecteurs lignes x_i à k dimensions pour toutes les valeurs de i et j , et β est le vecteur colonne des valeurs inconnues du paramètre. Nous utiliserons aussi la représentation matricielle pour les quantités obtenues sur échantillon. Donc, pour un échantillon de n éléments,

$$X' = (x_1', x_2', \dots, x_n'), \quad y' = (y_1, y_2, \dots, y_n).$$

Étant donné un échantillon de taille n et en considérant les x_i comme étant fixes, le meilleur estimateur (erreur quadratique moyenne minimale) de β est

$$\hat{\beta} = \left(\sum_{i=1}^n x_i' x_i \right)^{-1} \sum_{i=1}^n x_i' y_i = (X'X)^{-1}X'y. \quad (3.2)$$

où A est l'ensemble d'indices des éléments de l'échantillon et nous supposons, comme nous le ferons jusqu'à la fin, que la matrice à inverser est non singulière. Si les e_i n'obéissent pas à la loi de distribution normale, $\hat{\beta}$ est l'estimateur de la classe d'estimateurs linéaires non biaisés dont la variance est la plus faible. Nous pouvons écrire l'estimateur d'une combinaison linéaire de coefficients, disons

$$\theta^a = \sum_{j=1}^k \alpha_j \beta_j, \quad \text{sous la forme}$$

estimé. Särndal (1980), Wright (1983), ainsi que Särndal et Wright (1984) ont discuté des classes d'estimateurs par régression. L'ouvrage publié par Särndal, Swensson et Wretman (1992) contient une discussion détaillée de l'estimation par régression dont Mukhopadhyay (1993) a fait une revue.

Il a fallu attendre les années 1970 pour que l'on applique la méthode de régression aux données d'enquêtes générales à plusieurs variables et les années 1990, pour que l'application de la pondération par régression soit d'usage généralisé. Doane Agricultural Services Inc., appelé aujourd'hui Doane Marketing Research, est l'un des premiers organismes qui a utilisé les coefficients de régression pour la pondération. De 1971 à 1972, Doane a réalisé une étude d'audience en milieu agricole, à laquelle ont répondu 6 920 agriculteurs, sous la direction de M. John Wilkin. Les coefficients de pondération appliqués aux répondants ont été calculés par des méthodes de régression, en se servant de valeurs de contrôle provenant du U.S. Agricultural Census et du Department of Agriculture. Doane a offert un appui financier à la Iowa State University pour que celle-ci développe un programme de calcul de coefficients de pondération par régression. Lors de l'étude de Doane, pour s'assurer que les coefficients de pondération soient positifs, on a regroupé les observations pour lesquelles les coefficients de pondération étaient faibles et on leur a attribué un coefficient de pondération commun. Le regroupement s'est poursuivi jusqu'à ce que le coefficient commun soit positif. Plus tard, une version modifiée de la méthode de Huang et Fuller (1978) a été introduite dans les programmes informatiques pour garantir que les coefficients de pondération soient positifs. Doane applique les études de marché souscrites depuis 1972.

Statistique Canada a utilisé l'estimation par régression pour la première fois en 1988 pour l'Enquête sur la population active du Canada. En 1992, l'estimation par régression a été appliquée lors du *Recensement de la population du Canada* de 1991 afin de s'assurer que la somme pondérée des valeurs des variables recueillies au moyen du questionnaire détaillé (rempli par un échantillon systématique au 1/5 de l'ensemble des ménages du Canada) soit égale aux totaux connus au niveau des ménages et de la population recueillis lors du Recensement de 1991. À cet égard, consulter Bankier, Rathwell et Majkowski (1992) et Bankier, Houle et Luc (1997). L'estimateur par régression est également la composante essentielle du système généré par l'estimation (SGE) développé par Statistique Canada et utilisé pour un grand nombre d'enquêtes auprès des entreprises et d'enquêtes sociales depuis sa diffusion en 1992. La méthodologie qui sous-tend le système est décrite dans Estevao, Hidiroglou et Särndal (1995). Consulter aussi Hidiroglou, Särndal et Binder (1995). L'estimation par régression est maintenant utilisée pour construire les estimateurs composites pour l'Enquête sur la population active du Canada. À cet égard, consulter Singh, Kennedy et Wu

la régression dans le cas de l'échantillonnage d'enquête en s'appuyant fortement sur la théorie des modèles linéaires. Il a montré que le modèle linéaire ne doit pas nécessairement rester valable pour que l'estimateur par régression donne de bons résultats. Il a établi une expression $O(n^{-1})$ pour la variance, $O(n^{-2})$ pour la variance, et à l'égalité montrée que, dans le cas du modèle où la droite de régression passe par l'origine et la variance d'erreur est proportionnelle à x , l'estimateur par ratio est l'estimateur par les moindres carrés généralisés.

Durant les années 1950, l'estimation par régression a suscité un intérêt théorique qui a souvent pris la forme d'études du biais. À cet égard, consulter Mickey (1959), Brewer (1963) est l'un des premiers auteurs qui a considéré l'estimation linéaire lors de l'utilisation d'un modèle de superpopulation pour déterminer une méthode optimale. Il recherchait le plan d'expérience optimale pour l'estimateur par ratio et a examiné le conflit éventuel entre un plan optimal dans les conditions du modèle et un plan qui dépendrait moins de ce dernier. Consulter aussi Brewer (1979), Royall (1970) a soutenu que, dans le cas de l'utilisation de modèles, les propriétés conditionnelles importantes sont celles qui sont subordonnées à l'information auxiliaire contenue dans l'échantillon et que le plan d'échantillonnage devrait être choisi de façon à optimiser ces propriétés. Royall et ses collaborateurs, par exemple Royall et Cumberland (1981), ont étudié les propriétés conditionnelles des estimateurs par régression, découlant de l'., conditionnellement aux valeurs des variables auxiliaires obtenues dans l'échantillon.

Un grand nombre d'études effectuées durant les années 1970 et 1980 portaient sur la nature générale de l'estimateur par régression dans le contexte de l'échantillonnage et sur la mesure dans laquelle il est possible de rapprocher la méthode de prédiction par modèle et la perspective du plan d'échantillonnage. Fuller (1973, 1975) a énoncé les propriétés, dans le cas d'un grand échantillon, d'un vecteur de coefficients de régression calculé d'après un échantillon d'enquête. Isaki (1970) a étudié les estimateurs par régression et a publié des versions étoffées des résultats dans Isaki et Fuller (1982) et dans Fuller et Isaki (1981). Ces auteurs ont montré qu'un estimateur par régression construit dans les conditions d'un modèle est convergent par rapport au plan d'échantillonnage pour la moyenne de la population si le modèle contient certaines variables. Cassel, Särndal et Wretman (1976) ont tenu compte des éléments principaux du modèle et du plan d'échantillonnage pour construire l'estimateur et ont proposé l'expression « estimateurs par régression généralisée » pour les estimateurs de total convergents par rapport au plan d'échantillonnage de la forme

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (T_{\text{XN}} - \hat{Y}_{\text{HT}}) \hat{\beta},$$

où \hat{Y}_{HT} et \hat{Y}_{XHT} sont les estimateurs d'Horvitz-Thompson des totaux de y et x , respectivement, T_{XN} est le total connu de la population pour x et $\hat{\beta}$ est un coefficient de régression

une fois calculés, les coefficients de pondération plusieurs propriétés des estimateurs au cours de notre discussion. En nous appuyant sur un modèle, nous acceptons l'objectif classique consistant à minimiser l'erreur quadratique moyenne pour une classe d'estimateurs. Cette classe peut être celle des estimateurs linéaires qui sont dépourvus de biais dans les conditions du modèle, mais elle peut aussi faire l'objet de restrictions supplémentaires.

Les estimateurs qui ne varient en fonction ni de l'échelle ni de la localisation peuvent être utilisés dans des conditions générales. Mickey (1959) a proposé qu'on réserve l'expression « estimateur par régression » aux estimateurs linéaires qui ne varient en fonction ni de la localisation ni de l'échelle. Bien que nous ne puissions respecter strictement cette définition, nous appuyons la distinction entre les estimateurs invariants en fonction de la localisation et de l'échelle et ceux qui ne le sont pas. Selon nous, l'invariance selon la localisation est une propriété importante dans le cas des plans d'échantillonnage où l'unité visée par l'analyse est également une unité d'échantillonnage. Dans le cas des plans d'échantillonnage en grappes ou à deux degrés pour lesquels les coefficients de pondération sont construits pour l'invariance selon la localisation est moins importante.

Les modèles jouent un rôle considérable dans la construction des estimateurs par régression. Il est souhaitable que les propriétés des estimateurs restent bonnes même si la spécification du modèle n'est pas exacte. Par conséquent, les propriétés concernant la population finie visée par l'échantillonnage sont toutes aussi importantes que celles concernant le modèle.

Jessen (1942) et Cochran (1942) comptent parmi les premiers auteurs qui ont mentionné l'estimation par régression dans le cas de l'échantillonnage. Toutefois, le recours à la régression dans des contextes comparables a fort probablement eu lieu plus tôt. Cochran (1977, page 189) parlant en effet d'une régression en vue d'estimer la superficie d'une feuille réalisée par Watson (1937). Il est intéressant de noter que Jessen a utilisé la régression essentiellement dans le cas d'une estimation composite où la régression était destinée à améliorer les estimations calculées à deux points dans le temps, étant donné des échantillons ayant certains éléments communs à chacun de ces points. Cochran (1942) a énoncé la théorie de base de

2. CONTEXTE

Estimation par régression appliquée à l'échantillonnage

WAYNE A. FULLER

RÉSUMÉ

L'utilisation de la régression et des méthodes connexes est devenue courante dans le domaine de l'estimation à partir de données d'enquête. Nous passons en revue les propriétés fondamentales des estimateurs par régression, discutons de la mise en application de l'estimation par régression et étudions l'estimation de la variance des estimateurs par régression. Nous examinons aussi le rôle des modèles dans la construction des estimateurs par régression et l'utilisation de la régression en vue du rajustement pour tenir compte de la non-réponse.

MOTS-CLÉS : Information auxiliaire; calage; moindres carrés; convergence selon le plan; prédiction linéaire.

1. INTRODUCTION

Dans le domaine de l'échantillonnage, l'utilisation de renseignements sur la population étudiée est nécessaire lors de l'élaboration du plan d'échantillonnage et de l'estimation si l'on veut que les méthodes soient efficaces. Bien que ces deux activités soient intimement liées, puisque les estimateurs dépendent du plan d'échantillonnage, un traitement distinct leur est souvent réservé dans la documentation sur l'échantillonnage. Conformément à cette façon de faire, nous commençons par étudier l'estimation en traitant le plan d'échantillonnage tel que conçu. L'estimation consiste à combiner les renseignements disponibles sur la population et les données d'échantillon afin d'obtenir une bonne représentation des caractéristiques étudiées.

L'estimation par régression est l'une des méthodes importantes qui s'appliquent sur l'information démographique ou sur celle provenant d'un échantillon plus grand pour construire des estimateurs efficaces. Cette information, parfois appelée *information auxiliaire*, peut être utilisée lors de l'élaboration du plan d'échantillonnage ou ne pas être disponible à cette étape. Lors d'une enquête sur une population humaine, l'information provient souvent de sources officielles, telles que le recensement national. Des sources comparables peuvent fournir l'information auxiliaire pour d'autres catégories d'enquêtes. Par exemple, dans le cas d'une enquête sur l'utilisation des terres, les données sur la superficie totale, la superficie appartenant au gouvernement fédéral et la superficie immergée en permanence peuvent être extraites des archives nationales.

Trois situations peuvent se présenter en ce qui concerne l'information auxiliaire dont on dispose. Dans la première, on connaît les valeurs du vecteur x pour chaque élément de la population au moment de la sélection de l'échantillon. Dans ces conditions, on peut utiliser la variable auxiliaire pour établir le plan d'échantillonnage. Dans la deuxième situation, on connaît toutes les valeurs du vecteur x , mais on ne peut associer une valeur particulière à un élément particulier tant que l'on n'a pas

observé l'échantillon. Dans ces conditions, on ne peut utiliser l'information auxiliaire pour élaborer le plan d'échantillonnage, mais on dispose d'un éventail d'options d'estimation une fois que les observations sont disponibles. Par exemple, le recensement de la population peut fournir la répartition âge-sexe de la population, mais les organismes non gouvernementaux qui sélectionnent les échantillons ne disposent d'aucune liste de personnes et de leurs caractéristiques. Dans la troisième situation, on connaît uniquement la moyenne de la population pour x , ou la moyenne calculée d'après un grand échantillon. Dans ce cas, on ne peut utiliser l'information auxiliaire pour élaborer le plan d'échantillonnage et les méthodes d'estimation sont limitées. Par exemple, le U.S. Department of Agriculture pourrait diffuser une estimation du nombre total d'animaux d'un type particulier recensés sur les entreprises agricoles à une date donnée. Notre discussion se concentre sur cette dernière situation.

On peut aussi définir deux situations d'estimation. Dans un très petit nombre de paramètres, l'analyste est disposé à consacrer beaucoup d'énergie à l'analyse, il possède un modèle de population bien formulé et il est prêt à appuyer la méthode d'estimation en établissant le caractère raisonnable du modèle. Dans la deuxième situation, l'analyste s'attend à exécuter un grand nombre d'analyses portant sur un grand nombre de variables. Aucun modèle particulier n'est considéré adéquat pour toutes les variables. Le cas où le spécialiste de l'échantillonnage prépare un ensemble de données qui doivent être analysées par d'autres spécialistes est un exemple typique de cette deuxième situation. Comme une personne qui prépare l'ensemble de données ne sait pas quelles sont les variables analytiques, l'accent est mis sur l'utilisation d'estimateurs que l'on peut défendre en recourant à un nombre minimal de modèles. Les estimateurs par régression entrent dans la catégorie des estimateurs linéaires. Ces derniers présentent un avantage particulier dans le cas de l'échantillonnage, car,

MEMBRES DU COMITÉ DE SÉLECTION DE L'ARTICLE WASKBERG (2002-2003)

David A. Binder (Président), *Statistique Canada*
 J. Michael Brick, *Westat, Inc.*
 David R. Bellhouse, *University of Western Ontario*
 Paul Biemer, *Research Triangle Institute, U.S.A.*

Présidents précédents:

Graham Kalton (1999 - 2001)
 Chris Skinner (2001 - 2002)

Auteurs précédents:

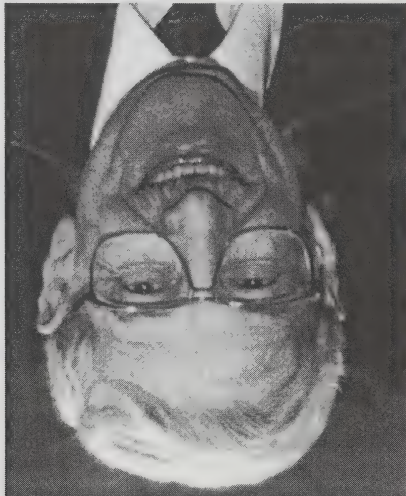
Gad Nathan (2001)

Nominations:

Les nominations d'individus à considérés comme auteurs ou les suggestions pour des sujets devraient être envoyés au président du comité, D.A. Binder, Statistique Canada, 3^e étage immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6, par courriel électronique: binderdav@statcan.ca ou par télécopieur (613) 951-5711. Les nominations et suggestions de sujets doivent être reçues au plus tard le 6 décembre 2002.

Série Waksberg d'articles sollicités

Le comité de rédaction de *Techniques d'enquête* a décidé de publier une série d'articles annuels sollicités en l'honneur de Joseph Waksberg, pour souligner sa contribution importante à la méthodologie d'enquête. Chaque année nous inviterons un spécialiste renommé de la recherche en sondages à rédiger un article consacré à la rétrospective et à l'examen de la situation courante d'un domaine important de la méthodologie d'enquête. L'auteur reçoit un prix monétaire grâce à une subvention offerte par Westat en reconnaissance de la contribution de Joe Waksberg durant les nombreuses années où il a travaillé pour l'entreprise. L'*American Statistical Association* est chargée de la gestion financière et administrative de la subvention. L'auteur de l'article est choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*.



JOSEPH WAKSBERG

ARTICLE SOLlicitÉ WAKSBERG 2002

Auteur : Wayne A. Fuller

Wayne A. Fuller est Professeur distingué émérite en statistique et économie à l'Iowa State University. Il a publié environ 100 articles dans plus de vingt revues et est auteur des livres *Introduction to Statistical Time Series* et *Measurement Error Models*. En tant que membre du Survey Group à l'Iowa State University, il a été principal responsable du développement de procédures d'estimation pour une grande enquête longitudinale appelée *U.S. National Resources Inventory*. Son intérêt pour la recherche en théorie des sondages comprend l'estimation par la régression, l'estimation pour petits domaines, l'imputation et l'échantillonnage à plusieurs phases. Il est présentement président du Comité consultatif des méthodes statistiques à Statistique Canada.

Dans leur article, Cahill et Chen élaborent une méthode d'exploitation des données provenant de plusieurs enquêtes et périodes de référence grâce à l'établissement des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie. Les estimations calculées d'après les données d'une enquête riche en variables explicatives sont étalonnées en fonction des renseignements fournis par une enquête dont l'horizon temporel est considérable. Cahill et Chen montrent comment appliquer la méthode à l'aide du module sur le congé de maternité du projet de microsimulation dynamique LifePaths de Statistique Canada.

Garrett et Chang examinent le problème de la population ne possédant pas de téléphone dans les enquêtes par téléphone au moyen de la composition aléatoire. D'après les échantillons de microdonnées à grande diffusion, la propension qu'un ménage ait un téléphone est estimée à l'aide d'une régression linéaire généralisée, propension dont on se sert pour l'estimation. Les biais et les variations asymptotiques sont présentés à la fois pour les estimateurs stratifiés à posteriori ainsi que non stratifiés à posteriori qui tiennent compte ou non de la propension estimée. Ces quatre estimateurs sont par la suite comparés dans le cadre d'une étude de simulation.

L'article de Tillie développe un estimateur qui permet d'éviter le problème des post-strates vides qui peut arriver avec l'estimateur post-stratifié usuel. L'idée consiste à utiliser un estimateur pondéré conditionnellement et à conditionner sur les rangs dans la population d'une variable auxiliaire connue pour toutes les unités de cette population. De cette façon, les tailles des post-strates sont fixées dans l'échantillon et sont aléatoires dans la population. Ensuite, on calcule la moyenne d'estimateurs pondérés conditionnellement pour obtenir plus de stabilité. L'estimateur obtenu est calé sur la répartition, linéaire et exactement sans biais. On montre au moyen d'une étude de simulation que l'estimateur proposé est plus robuste que l'estimateur par la régression généralisée quand la relation entre la variable d'intérêt et la variable auxiliaire n'est pas linéaire. On propose finalement un estimateur approché de la variance valide au moyen de simulations.

Dans leur article, Shao et Butani traitent du problème de l'estimation des variances dans le cas des estimateurs d'enquête imputés. Ils montrent que les variances qui résultent peuvent être estimées en deux volets : il s'agit d'abord de procéder à une estimation à l'aide d'une méthode de regroupement en demi-échantillons groupé qui incorpore des ajustements de sorte que l'on tienne compte de l'imputation. Comme l'estimation du deuxième volet peut donner lieu à de nombreux déviations, Shao et Butani proposent que l'on procède à une correction de la méthode de regroupement en demi-échantillons groupé qui donnerait lieu à des estimations des variances à peu près sans biais.

Dans son article, Cohen décrit une méthode pour mettre en œuvre la méthode jackknife de Rao et Shao concernant l'estimation des variances de manière à tenir compte de l'imputation de valeurs au moyen des poids de rééchantillonnage. La méthode de Rao et Shao suppose le calcul, pour chaque méthode de rééchantillonnage, des valeurs corrigées des données imputées. La méthode peut être utilisée soit avec l'imputation de la moyenne ou soit avec l'imputation par la méthode du hot deck. Pour appliquer la méthode de Cohen, il faut ajouter des rangées supplémentaires au fichier de poids de rééchantillonnage. Pour chaque valeur imputée, on ajoute une rangée supplémentaire pour chaque répondant faisant partie de la même classe d'imputation.

Dans le dernier article du présent numéro, Valliant étudie plusieurs estimateurs des variances pour l'estimateur de régression généralisée (GREG). Ce qui nous intéresse c'est de trouver des estimateurs de la variance qui, sous réserve de certaines conditions, ne renferment à peu près pas de biais tant du point de vue de la variance due au plan de sondage que du point de vue de la variance due au modèle même si le modèle qui sous-tend l'estimateur GREG renferme un paramètre de variance qui est faux. Soulignons qu'une caractéristique importante de ces estimateurs robustes est l'ajustement des carrés des résidus au moyen de facteurs analogues aux effets leviers utilisés en analyse par régression classique. Il semble que l'estimateur jackknife avec suppression d'une unité inclut les ajustements pour tenir compte des effets leviers et qu'il soit un bon choix du point de vue tant de la variance due au plan de sondage que de celle due au modèle. Une étude de simulation montre que ces estimateurs de la variance sont caractérisés par un biais faible et produisent des intervalles de confiance dont le taux de couverture est quasi normal.

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient le deuxième d'une série d'articles annuels sollicités et publiés en l'honneur de Joseph Waksberg. Une brève description de la série et une courte biographie de Joseph Waksberg ont été présentées dans le numéro de juin 2001. L'auteur de l'article de 2002 de la série Waksberg est Wayne Fuller. J'aimerais remercier les membres du Comité, Graham Kalton (président), Chris Skinner, David Binder et Paul Biemer d'avoir choisi un statisticien aussi prestigieux, qui a largement contribué à de nombreux aspects de la théorie et de la pratique statistiques, comme auteur du deuxième article de la série Waksberg.

Dans son article intitulé *Estimation par régression appliquées à l'échantillonnage*, Wayne Fuller présente un aperçu général des faits historiques et des faits nouveaux en ce qui a trait à l'utilisation des modèles de régression dans les enquêtes pour l'estimation, la calibration des poids et le facteur de compensation de la non-réponse. Après une brève introduction et un aperçu historique, il aborde la question de l'utilisation des modèles de régression pour l'estimation dans le cadre d'enquêtes complexes du point de vue de la variance due au plan de sondage. Il enchaine avec une étude du point de vue de la variance due au modèle. Il aborde également l'utilisation des modèles de régression pour les données multivariées, les techniques à employer quand on dispose de variables auxiliaires pour chaque unité de la population et la régression de manière à tenir compte des effets de la non-réponse dans les enquêtes. Enfin, en exposant certains des aspects pratiques des applications, l'auteur complète cet aperçu éclairant d'un important domaine d'inférence à partir de données d'enquête auquel Wayne Fuller a lui-même fait de nombreux apports considérables.

Ce numéro renferme également une section spéciale intitulée *L'héritage de Leslie Kish* qui comprend quatre articles, dont un rédigé par Leslie Kish lui-même qui fait part de ses dernières pensées sur la combinaison des échantillons et des enquêtes. Deux autres articles traitent de la mise en œuvre de l'idée de Leslie Kish concernant les recensements consécutifs. Ces deux articles ont aussi été présentés dans le cadre du Symposium de 2001 de Statistique Canada lors d'une séance spéciale intitulée *L'héritage de Leslie Kish*.

Le premier article de la section spéciale, par Graham Kalton, présente un aperçu inspirant de l'apport de Kish à de nombreux secteurs de la statistique. Bon nombre des problèmes sur lesquels Kish a travaillé sont placés dans un contexte historique, et leur importance pratique est soulignée. L'article de Kish présente des problématiques sur lesquelles il travaillait au moment de son décès survenu en octobre 2000. Je remercie Graham Kalton et Jack Gambino des corrections qu'ils ont apportées à l'article, bien que celui-ci soit reproduit en grande partie tel qu'il l'était au moment du décès de Kish. Dans l'article, il prétendait que, tout comme la statistique représentait un nouveau paradigme dans la méthode scientifique et que l'échantillonnage nécessitait un nouveau paradigme en statistiques, les échantillons consécutifs et les enquêtes couvrant plusieurs populations nécessitent de nouveaux paradigmes en ce qui a trait aux méthodes d'enquête. Nous ne pouvons que formuler des hypothèses sur ce que l'article final aurait été si Kish vivait toujours.

Alexander décrit l'*American Community Survey*, que doit lancer le U.S. Census Bureau dans les prochaines années à titre de remplacement du questionnaire détaillé du recensement décennal. Il s'agit là d'une enquête très vaste qui se fonde en grande partie sur le concept d'échantillons et de recensements consécutifs que Kish a lancé il y plus de 20 ans. Cet article traite des concepts, de la base de sondage, du plan de sondage, du cumul d'échantillons et de la pondération.

Le dernier article de la section spéciale, rédigé par Durr et Dumais, décrit le nouveau recensement consécutif lancé en France dans le but de remplacer le recensement plus traditionnel. Dans le cadre de ce recensement consécutif, chaque petite commune fera l'objet d'une enquête une fois au cours d'une période de cinq ans; les grandes communes seront divisées en cinq groupes de renouvellement dont chacun fera l'objet d'une enquête une fois tous les cinq ans. Dans l'article, on traite des objectifs, des procédures quant au plan de sondage et à l'estimation qui s'appliquent aux recensements consécutifs.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 28, numéro 1, juin 2002

TABLE DES MATIÈRES

Dans ce numéro	1
Article Sollicité Waksberg	
W. FULLER	5
Estimation par régression appliquée à l'échantillonnage	
Section spéciale "En souvenir de Leslie Kish"	
GRAHAM KALTON	
L'influence de Leslie Kish sur la statistique d'enquête	27
LESLIE KISH	
Nouveaux paradigmes (modèles) pour l'échantillonnage probabiliste	33
CHARLES H. ALEXANDER	
Les échantillons successifs de Leslie Kish et l'American Community Survey	39
JEAN-MICHEL DURR et JEAN DUMAIS	
La rénovation du recensement français	47
Articles Réguliers	
IAN CAHILL et EDWARD J. CHEN	
Estimation des paramètres estimés des modèles logit de choix binaire et des modèles semiparamétriques de survie	55
STEVEN T. GARREN et TED C. CHANG	
Estimation améliorée des ratios dans le cas des enquêtes téléphoniques avec correction pour la non-couverture	67
YVES TILLE	
Estimation sans biais par calage sur la répartition dans les plans simples sans remise	83
JUN SHAO et SHAIL BUTANI	
Estimation de variance dans le cadre de la « Current Employment Survey »	93
MICHAEL P. COHEN	
Application de l'estimation de la variance selon Rao-Shao en utilisant des poids de rééchantillonnage	103
RICHARD VALLIANT	
Estimation de la variance de l'estimateur de régression généralisée	109

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

- Président** G. J. Brackstone
Membres D. A. Binder
G. J. C. Hole
C. Patrick
R. Platek (Ancien président)

COMITÉ DE RÉDACTION

- Rédacteur en chef** M. P. Singh, *Statistique Canada*
- Rédacteurs associés**

- D. R. Bellhouse, *University of Western Ontario*
D. A. Binder, *Statistique Canada*
J. M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W. A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M. A. Hidiroglou, *Statistique Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*
G. Nathan, *Hebrew University, Israel*

- Rédacteurs adjoints** J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *Statistique Canada*

- D. Norris, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
J. N. K. Rao, *Carleton University*
T. J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
F. J. Schuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C. J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K. M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découplant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M. P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de *Techniques d'enquête* (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Juin 2002

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, par quelque support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2002

Publication autorisée par le ministre
responsable de Statistique Canada

JUN 2002 • VOLUME 28 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

•

VOLUME 28

•

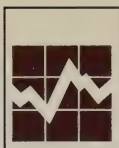
JUIN 2002

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE





SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2002

•
VOLUME 28

•
NUMBER 2





SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2002 • VOLUME 28 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2003

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

January 2003

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Past Chairman)

E. Rancourt (Production Manager)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*
G. Nathan, *Hebrew University, Israel*

D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY
A Journal Published by Statistics Canada
Volume 28, Number 2, December 2002

CONTENTS

In This Issue	115
GORDON J. BRACKSTONE Strategies and Approaches for Small Area Statistics	117
DENNIS TREWIN The Importance of a Quality Culture	125
YVES THIBAUDEAU Model Explicit Item Imputation for Demographic Categories	135
BALGOBIN NANDRAM, GEUNSHIK HAN and JAI WON CHOI A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas	145
JAY STEWART Assessing the Bias Associated with Alternative Contact Strategies in Telephone Time-Use Surveys	157
ROBERT M. BELL and DANIEL F. MCCAFFREY Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples	169
MONROE G. SIRKEN Design Effects of Sampling Frames in Establishments Survey	183
LOUIS-PAUL RIVEST A Generalization of the Lavallée and Hidioglou Algorithm for Stratification in Business Surveys	191
WILSON LU and RANDY R. SITTER Multi-way Stratification by Linear Programming Made Practical	199
ROBBERT H. RENSSSEN and GERARD H. MARTINUS On the Use of Generalized Inverse Matrices in Sampling Theory	209
Acknowledgements	213



In This Issue

This issue of *Survey Methodology* includes papers on a variety of topics including overviews of small area statistics and data quality in statistical offices, survey nonresponse and imputation, survey design, data collection and estimation.

In the first paper of this issue, Brackstone identifies strategies and approaches for the development of small area statistics programs in national statistical offices. The topic of small area estimation will be covered by a number of papers in a special section in the June 2003 issue of *Survey Methodology*. The paper first considers the crucial role of censuses, and discusses issues related to their usefulness for small area statistics. Other potential sources of small area statistics include administrative files and sample surveys, either on their own or combined with census data to provide estimates for the intercensal period or for characteristics not directly covered by the census. Rolling censuses are also discussed, as well as the unique challenges in producing small area business and environmental statistics. Finally, issues of organization of national statistical offices for production and dissemination of small area statistics are considered.

Trewin reviews the practices and approaches used to maintain high quality of output from a national statistical office. Important ingredients include good relations with respondents, skilled and motivated staff, sound statistical and operational methods, and relevance of statistical programs. Current challenges include increasing the use of administrative data sources, effective use of the internet for both collection and dissemination, maintaining knowledge and skills as staff leave, and handling increasing user expectations. This paper is based on a talk presented as the keynote address at Statistics Canada's Symposium 2001.

Thibaudeau presents an innovative approach to the imputation of demographic characteristics in a large scale survey or Census. Instead of relying on the usual approach of either the closest complete record in the processing stream or constructing imputation groups, Thibaudeau proposes a compromise method which uses maximum likelihood estimation based on the conditional probabilities. This approach seeks to create groups that are close in order and in geography to the imputed record. He also presents an interesting Bayesian approach to evaluating the method.

Nandram, Han and Choi consider the problem of analyzing multinomial nonignorable non-response data from small areas in the framework of Bayesian inference. This paper extends some earlier work by Stasny by assuming a Dirichlet prior underlying the multinomial probabilities and using a prior distribution on the hyperparameters. The authors apply this model to Body Mass Index data from a complex survey design.

In the Stewart paper, the possible biases introduced by different contact strategies in telephone time-use surveys are investigated. Two contact strategies, convenient-day scheduling, where the designated reference day changes with the contact day, and designated-day scheduling, where the reference day remains fixed, are discussed and compared through simulation studies.

Bell and McCaffrey consider the problem of unbiasedly estimating the variance of coefficients of linear regressions from multi-stage survey data when only a small number of Primary Sampling Units (PSUs) are sampled. After investigating situations where the bias of the linearization variance estimator can be large, a bias reduced linearization variance estimator is proposed. In addition, a Satterthwaite approximation is used to determine the degrees of freedom to be used for tests and confidence intervals in conjunction with the bias reduced linearization variance estimator.

Sirken considers estimation of the volume of transactions that a population of establishments has with a population of households. An approach based on indirect sampling of establishments through the households that they have transactions with is compared to the more typical approach based on direct pps sampling of establishments. Estimators and expressions for the variances are derived and compared for the two methods. Situations where one approach or the other is preferable are explored.

Rivest considers the problem of identifying stratum boundaries. The commonly used Lavallée-Hidiroglou algorithm assumes that the values of the study variable are available and are used in the determination of optimal stratum bounds. In his paper, Rivest relaxes this assumption and modifies the Lavallée-Hidiroglou algorithm to account for a discrepancy between the stratification variable and the study variable through the use of models that link these two variables together. These models are then incorporated into the Lavallée-Hidiroglou algorithm.

In the Lu and Sitter paper, the problem of the sample size being smaller or only slightly larger than the total number of strata is considered. Consequently, conventional methods of sample allocation to strata may not be applicable. One solution for this problem is to use a linear programming technique to minimize the expected lack of desirability of the samples subject to a constraint of expected proportional allocation (EPA). However, as the number of strata increases this solution rapidly becomes expensive in terms of magnitude of computation. In the proposed approach, the amount of computation is reduced substantially at the small cost of approximate EPA for strict EPA.

Renssen and Martinus explore the use of generalized inverse matrices in survey sampling. After reviewing the properties of generalized inverses, they consider the generalized regression estimator when the set of regressors is not of full rank, and they set out a regularity condition under which the estimator is invariant to the choice of generalized inverse. They then present an algorithm for calculating the regression weights, and briefly discuss weighting in the Dutch Labour Force Survey.

M.P. Singh

Strategies and Approaches for Small Area Statistics

GORDON J. BRACKSTONE¹

ABSTRACT

National statistical offices are often called upon to produce statistics for small geographic areas, in addition to their primary responsibility for measuring the condition of the country as a whole and its major subdivisions. This task presents challenges that are different from those faced in statistical programs aiming primarily at national or provincial statistics. This paper examines these challenges and identifies strategies and approaches for the development of programs of small area statistics. The important foundation of a census of population, as well as the primary role of a consistent geographic infrastructure, are emphasized. Potential sources and methods for the production of small area data in the social, economic and environmental fields are examined. Some organizational and dissemination issues are also discussed.

KEY WORDS: Small area statistics; Census; Geography.

1. INTRODUCTION

The mandate of most national statistical offices (NSO) focuses on the monitoring of social, economic, and environmental conditions at the national level, and for the major administrative units (provinces, states, major metropolitan areas) within the country. However, the demand for data at lower geographic levels is always present, especially from local governments and from businesses needing to make investment, marketing, and location decisions that depend on knowledge of local areas. We will use the term "small area statistics" to mean statistics for areas below the level of state, province, or major metropolitan areas – a broad spectrum of areas from large towns, through urban neighbourhoods, to rural villages. In some circles the term "small areas" is used more broadly to refer to any small sub-group or domain of the population, but here we are talking strictly about small geographic areas.

The extent of an NSO's responsibility for small area statistics depends on the division of governmental responsibilities within a country. For example, in some countries local governments are the creation of provinces and the responsibility for supporting their statistical needs may rest with provincial governments. But in many countries, whatever the formal division of powers, it is, *de facto*, the NSO that is expected to respond to the need for small area statistics, either within its own resources or in cooperation with other levels of government. At the very least, it is the NSO that must set the standards and framework for small area data if these are not to become a mishmash of uneven and overlapping statistics incomparable across the country.

With limited budgets an NSO is faced with the difficult trade-off between investment in national statistics and provision of small area detail. How should it choose between covering more subject areas, or existing subject areas in more detail, at the national and provincial levels, and, on the other hand, providing more small area detail for

subject areas it is already covering nationally? There is no formula for resolving this problem. The balance struck in any country will be largely a function of national needs, relative powers, and historical tradition, with perhaps some statistical considerations on the margin. Nevertheless, there is a series of measures and approaches that a NSO can consider to maximize the degree to which it can satisfy demands for small area statistics within a limited budget.

Four potential sources of small area statistical data either individually or in combination, account for most production of small area data by statistical agencies. Censuses or complete enumerations of populations are the traditional source. Administrative records, including national registers, that cover all, or almost all, of a defined population are in many respects equivalent to a census. National sample surveys are rarely large enough to produce small area data directly but they do represent a valuable current source of information that can be used, under certain assumptions and in combination with other sources, to produce small area data. And finally, local studies focused on particular small areas will produce small area data, but not for complete sets of small areas. Sources such as satellite imaging or aerial photography can be thought of as censuses or local studies depending on their coverage.

In this paper we first review the important role of the Census of Population, with or without a population register, in the provision of small area socio-economic data (Section 2), and then emphasise the fundamental role of an up-to-date geographic infrastructure to support any production of small area statistics, including especially the census of population (Section 3). We then examine approaches to providing small area data on individuals and families between censuses (Section 4), on business activities (Section 5), and on environmental issues (Section 6). We conclude with some general observations about the dissemination of small area statistics and the management of small area statistics within an NSO.

¹ Gordon J. Brackstone, Informatics and Methodology Field, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: bragcor@statcan.ca.

2. CENSUS OF POPULATION

The census of population, in most countries, plays the central role in the provision of small area data about people, families and households. Based on a complete enumeration of the population (at least for basic characteristics), its estimates are free of the sampling error that limits the ability of sample surveys to produce small area estimates. Provided the individual households are geographically coded to a fine level (*e.g.*, a block or block face), direct tabulations of households can produce statistical aggregates for any geographic area that can be defined, or approximated, in terms of the lowest level of geographic coding.

However, censuses have their drawbacks. They are costly, and therefore they are infrequent. Data from the last census may provide a poor representation of a small area that is undergoing rapid development. In many countries, sampling is utilized in the census for many of the questions. While this introduces sampling error into estimates from the census, these samples are still huge compared with those in a typical sample survey. Furthermore, the samples are typically spread through every enumeration area of the country, so the ability to produce small area estimates is maintained, even though the small areas will need to be somewhat larger than in a true census.

Potentially more serious, with respect to accuracy, are nonsampling errors such as coverage error and response bias. Most censuses miss some people, or count some people twice, and it has been repeatedly shown that those miscounted are generally not typical of the population as a whole. Census estimates may therefore be biased against certain sub-groups of the population. If these subgroups (*e.g.*, certain immigrant groups) tend to be geographically clustered, this can have a serious impact on estimates for some small areas. Response bias arises if a census question is systematically misunderstood by many respondents. Both small area and large area estimates would be affected by such errors.

Countries that maintain a population register have the potential to produce census-like data for small areas more frequently than the traditional 5-10 year cycles of a census. Up-to-date residence registration is clearly a requirement for accurate small area data from such registers. The breadth of data available from a register system may be less than that available through a conventional census, since the former is limited to the characteristics maintained in linkable administrative registers. In some countries the population register may be used as the basis for a census that collects the necessary additional characteristics not available within existing registers Redfern (1989) provides a useful description of practices within Europe in this regard.

Since the Census has the potential to produce estimates for very small areas, rules to protect against direct or residual disclosure of individual data have to be in place. These can include imposing a minimum population on areas for which data will be released, random perturbation of

data, suppression of data, or other techniques (Jabine (1993), Zayatz, Steel and Rowland (2000)). NSOs have also to be concerned about privacy issues arising with the publication of small area census data that, while not disclosing any individual responses, do reveal dominant characteristics of an area (*e.g.*, that 90% of the families received unemployment benefits). Such findings cannot be withheld, but they can be selected and presented with sensitivity.

Though a census, with or without a population register, is a source of direct small area data as of census day, the value of such data declines as time passes. However, the role of census data in the provision of small area statistics goes well beyond the direct use of the results from each periodic enumeration. Inter-censally, census data may be used as a benchmark, a sampling frame, or as auxiliary information to be used with other sources of data that are available between censuses. These usages are pursued in section 4. An innovative alternative to the traditional census is described in Section 4.4.

3. GEOGRAPHIC INFRASTRUCTURE

To enable a national census to produce accurate data for small areas, a geographic infrastructure of boundaries and mapping capacity covering the whole country is a prerequisite. Such an infrastructure requires that each dwelling be associated with a precise geographic location on the ground, where the degree of precision determines the fineness with which small areas can be defined. Though modern global positioning technology makes it possible to pinpoint each dwelling to a specific pair of coordinates, it is usually sufficient for statistical purposes to associate each dwelling in an urban area with a block face (*i.e.*, one side of a street between two intersections), or a building in the case of high-rise buildings. In rural areas, the chosen degree of precision will depend on local administrative and natural boundaries, though maximum flexibility is preserved by using precise coordinates for each dwelling.

While necessary for a census, a geographic infrastructure is equally required for the provision of small area statistics from other sources. Essentially each data point, from whatever source, has to be associated with a geographic location at a level detailed enough to allow aggregation into any small areas of statistical interest. For example, if the data source is an administrative register, or a business register, the address in each record must be convertible into a pair of geographic coordinates, or at least into a small area within which the address falls. Since administrative registers often use mailing addresses, a file that converts postal codes into geographic locations is a valuable tool in the development of small area data.

The availability of an accurate up-to-date geographic infrastructure, whether maintained by the NSO or obtained from outside, is essential if a program of small area statistics is to have flexibility in the choice of areas for which statistics are produced.

4. SMALL AREA STATISTICS ON PERSONS AND HOUSEHOLDS – BETWEEN CENSUSES

We turn now to the issue of producing small area data for persons or households inter-censally. Clearly the existence of a current population register makes a fundamental difference to what is possible, and how it can be done. We will confine ourselves to the case where no regularly updated population register exists.

In such circumstances, there are three main classes of approach. The first is to utilize census-like files that come from administrative systems and purport to cover the whole of a well-defined population. The second is to exploit sample survey data and, through additional model assumptions, produce estimates for smaller (though still not very small) areas than is possible through direct survey estimation. The third category is the combination of one or both of these first two approaches with the use of data from the most recent census. In the following paragraphs we review some of the characteristics of these approaches.

4.1 Administrative Files

An example of an administrative file with small area statistical potential is the annual file of individual income tax returns. Other examples, with narrower population definitions, might be drivers' licences, employment insurance recipients, or health insurance records. In the case of tax data, if each record contains a residential address that can be associated with a geographic point or small area, then data can be tabulated directly for small areas, with due regard for confidentiality (as with census data). The characteristics available would generally be restricted to demographic and income variables, and the coverage would be limited to taxfilers. Nevertheless, such a file represents a rich source of annual data for quite small areas. Population coverage can be improved through the imputation of dependents "claimed" on the tax record. In Canada, the coverage of such imputed files is approaching that of the census as coverage increases among low income earners who need to file tax returns to obtain social assistance benefits.

With administrative data in general, the statistician has to take what is available (though some influence on content may be possible in the longer term), reconcile any differences in concepts, definition or coverage between the administrative file and the statistical objectives, and assess any issues of reporting or coding accuracy in the records. Subject to these precautions, administrative data can provide a geographically rich potential source of small area data (Brackstone 1987).

4.2 Sample Survey Data

The problem with sample survey data as a source of small area statistics is sample size. There are frequently insufficient sample cases in the small area to allow a reliable direct estimate to be produced, and sometimes none

at all. In large national sample surveys it may be possible to devise sampling strategies that ensure an acceptable level of precision for planned small areas, such as sub-provincial regions, without significantly degrading the reliability of estimates at higher levels (Singh, Gambino and Mantel 1994). But for smaller areas, or for areas of similar size not taken into account during design, reliable estimation will not be possible. Larger samples help, and may allow direct estimation for some of the larger small areas, but budgets usually constrain this approach as a general solution. If no other data sources are available, statisticians can only resort to model-based methods which involve making assumptions about how data for a small area relate to other data. These methods are often described as "borrowing strength" since they borrow information from elsewhere in the sample survey to augment the number of units that contribute to the estimate for a given small area. The borrowing can be from other time periods, from sample units outside the given small area, or from other variables measured on the same sample unit. Some examples follow. Most of these examples will allow some expansion of the range of small area estimates that can be produced from sample surveys with relatively large samples. They cannot magically convert small sample surveys into rich sources of small area data.

1. In a monthly survey, it may be possible to combine data for a small area over a period of consecutive months to produce direct estimates of a multi-month moving average for the area. For example, quarterly estimates may be possible where monthly ones were not.
2. One may be ready to assume that means or proportions estimated for a larger area apply equally to a smaller component area within it. If the size of the small area is known, an estimate can be obtained by multiplying by the assumed mean or proportion. This assumption may be more realistically made within subgroups of the population (*e.g.*, age groups), rather than for the population as a whole. In this case, if the size of each sub-group is known for the small area, a synthetic estimator can be built up by multiplying the sizes by the assumed means and aggregating.
3. If additional related variables are available from the survey, more elaborate models may be set up relating the variable being estimated to these auxiliary variables. The parameters of the model may be estimated at a higher geographic level where there is sufficient sample to estimate them reliably. The model is then applied with the estimated parameters to the data for the given small area.

All of these approaches suffer from the lack of reliable baseline data for each small area. If such data are available,

for example from a recent census or from administrative records, then the data may be used in combination to produce more reliable estimates than from either source alone.

4.3 Combined Sources

Methods that combine census or administrative information from the recent past with current sample survey data are borrowing strength from outside the survey. They still require model assumptions. However, these can often be weaker (since they involve assumptions about change from the benchmark, rather than about absolute levels of each small area) and so more acceptable, or more plausible, than in the case of sample survey data alone.

A wide variety of estimation methods (which we won't attempt to describe here) have been developed to handle this situation. Some of these methods can be thought of as estimating change since the most recent benchmark, others as distributing reliable current sample survey estimates among component small areas based on benchmark data, and yet others as recalibrating old benchmark figures to new current estimates. In essence, they all involve some kind of balancing of three kinds of estimates: (a) high variance but unbiased direct current survey estimates for the small area in question; (b) low variance current survey estimates for some surrounding or comparable larger area; and (c) census-type estimates for the same small area from recent administrative data, or a past census, which may contain unknown bias due to the source and the time lag. Any available auxiliary data can be incorporated to improve the accuracy of each component estimate. The way in which these three types of estimates are combined is determined by the choice of model and model parameters.

In summary, the methods of this and the previous section essentially reduce variance by making use of more data, but at the expense of introducing potential bias due to model assumptions that will never be exactly correct. It is very important to analyse the performance of these methods before their use, for example by carrying out the estimation process in a census year when direct estimates are available for comparison, and periodically thereafter. Model checking is becoming an area of increased research activity (Bayarri and Berger 2000). For more detailed descriptions of available methods in this class see, for example, Purcell and Kish (1979); Fay and Herriott (1979); Ghosh and Rao (1994); Singh *et al.* (1994); Schaible (1996); Rao (1999) and Gambino and Dick (2000).

4.4 Rolling Censuses

An innovative alternative to the census is being investigated in at least two countries. The method of producing small area data based on a large rolling sample has long been advocated by Leslie Kish as an alternative to the traditional census (Kish 1990, 1998). The sample survey "rolls" in the sense that over a long period (*e.g.*, a decade) each of the smallest areas for which estimates are required

would be included once in the sample so as to provide a direct estimate for that area once each period. Successively larger areas (aggregates of the smallest areas) would be represented more often in the sample, allowing either more reliable or more frequent estimates for those areas. For even larger areas, including provinces and the whole country, the accumulated sample would be sufficient to provide reliable annual, or more frequent, estimates at certain levels of detail. The approach may be considered with or without a periodic census to collect basic demographic data against which to calibrate the inter-censal survey estimates.

The rolling census avoids the need for the assumption of models, but presumes that unbiased estimates of multi-year averages, or asynchronous estimates for different areas of the country, are satisfactory alternatives to the simultaneous point-in-time estimates of the traditional census. Relative cost is also a key factor, especially in the situation where a basic census is also carried out. On the other hand, by producing reliable annual estimates for many of the larger areas, and with much of the content detail of a census, this approach could effectively address the issue that census estimates can be up to 12 years old before the next ones appear. It also responds to mounting concerns over increasing difficulties and costs associated with the conduct of a traditional census.

This approach is being tested in the United States under the name of the American Community Survey (Alexander 1999, 2002) and in France where it is referred to as the "recensement continu" (Isnard 1999; Durr and Dumais 2002).

5. BUSINESS STATISTICS

The problems of producing small area data for businesses are different in many important respects from those encountered for data on persons or households.

Whereas the association of each individual with a "usual place of residence" is, for the vast majority of the population, a fairly clear and unambiguous concept (though perhaps becoming less clear with the growth of second residences, the incidence of prolonged absences away from the snow, and more flexible living arrangements), for businesses the question of where, geographically, to attribute various characteristics of a business is less clear in many situations. For single establishment businesses where all the activity takes place in a single location there is no conceptual problem, though there may still be a practical problem if the source of information is an administrative file that provides, say, an accountant's address rather than the place of business. For some variables, such as employment, there may be no major conceptual problem even for larger businesses (except perhaps for those working in the transportation industry, or certain service industries). However, for variables such as revenues and profits there can be real questions about how these should

be allocated geographically in multi-establishment businesses. The larger the geographic area the smaller the problem – location within a province doesn't matter if one is only interested in provincial totals. But, in general, geographic attribution rules have to be determined before small area estimates for business activity can be considered, and for some aspects of business activity small area estimates may not make conceptual sense.

While for household surveys the main obstacle to the production of small area estimates is sample size, for business surveys considerations of confidentiality usually constitute the major barrier. The smaller the area, the greater the chance that a particular industry will be dominated by one or a few major companies, thus precluding the provision of estimates for that area due to disclosure risk. Methods for checking statistical output on businesses to recognize potential disclosure risks are fairly well developed (Federal Committee on Statistical Methodology 1994) but require constant attention on the part of the NSO. The confidentiality problem is less of an issue in those industries characterized by small units – which may be the same industries in which the conceptual problems of the previous paragraph are not so severe. In those industries, considerations of sample size may indeed be the limiting factor, in which case the families of methods described in the previous section are available.

A third area of contrast with data on individuals, at least for countries that do not maintain a population register, is the existence of a relatively up-to-date list frame of businesses. This not only provides a base for sampling and a source of some auxiliary data for estimation, but also constitutes a potential source of direct estimates of business demography, at least annually. In many countries the currency of the business register is maintained by receiving transactions from the business tax system, which itself provides an annual census-like source of administrative data on business activity. However, use of tax data still requires careful consideration of the conceptual, geographical and confidentiality issues raised above.

6. ENVIRONMENT STATISTICS

Environment statistics provide yet different challenges for the production of small area statistics. While some environmental issues are national or even global in scope, many are by their nature local. Many sources of pollution are typically localized with their impacts being felt most severely in the neighbourhood of a plant or accident. The socio-economic impacts of broader environmental problems (e.g., loss of fish stocks) are frequently felt in small and often isolated resource-based communities.

Some environment data are collected from households or individuals (e.g., recycling practices, fuel use) and their potential as a source of small area data is subject to the considerations already described in section 4. Other

environment data (e.g., waste generation, environmental protection expenditures, use of natural resources) come from businesses and would be governed by the considerations of Section 5. However, a great deal of environment data is obtained from physical surveys (e.g., geological, physiographic, hydrographic), from instrument measurement (e.g., temperature, air quality, water quality, ozone layer thickness), and from direct observation (e.g., land use). Different considerations govern the relation of these data sources to small area data.

Because environment data are no respecters of administrative boundaries, the need for a flexible geographic infrastructure, emphasised in Section 3, is especially important here. Small area geographic identification is needed to regroup data to geographical units that are more suitable for environmental analysis. For example, the production of waste attributable to a certain type of agricultural activity might be aggregated for all of the producers within a river basin. Environmental geographic units are either pre-defined (ecozones, drainage basins) or dictated by special events (areas covered with different thicknesses of ice, land areas flooded by heavy rains or spring thaws). In some cases, the area studied could be a very small site such as a park.

Physical quantity or quality data can be difficult to aggregate or summarize. In some cases, point source data such as air quality measures cannot be considered representative of any larger geographic unit. Water quality may be summarized or compared by using an indicator, such as the number of days beaches are open for swimming, but not simply as an aggregate or average of water quality readings. For many measures, the focus of interest may be on change over time rather than small area comparisons. In other cases, sampling and estimation techniques may need to make use of spatial analysis techniques such as contouring or interpolation.

The privacy and confidentiality concerns associated with environment data depend on their source. Data collected from households or businesses, even if they involve physical measurements, are protected by the same confidentiality rules as other data from those sources. Direct measurements of the stock of natural resources or the quality of the environment do not raise these concerns. Cartographic representation of spatial patterns may be one way to overcome some of the analytical frustrations of data suppression for small areas. Choropleth maps (maps which show the distribution of variables or characteristics by using colour or shading for ranges of the distribution) can explicitly represent the ranges implicit in rows or columns that would be suppressed in a published table.

Cross-border pollutant flows and their global effects make physical environment data an international issue. Cooperation between neighbouring countries is necessary to ensure that national boundaries do not impede analysis of the impact of physical processes that recognize no such boundaries.

In summary, the small area dimension is particularly important for environment data, not only because a locality is frequently the point of interest, but also because data must often be reaggregated to geographic areas more appropriate for environmental analysis such as ecozones or watersheds.

7. ORGANIZATION AND DISSEMINATION ISSUES

Most NSOs are organized by subject-matter area. The production of small area estimates cuts across subject-matter areas, but requires support from Geography staff for geographic infrastructure, from Methodology staff for estimation and evaluation methods, and perhaps from other staff for analyzing and packaging data across subject areas. The question of how to organize small area estimation within an NSO therefore arises.

Requiring subject-matter areas to manage small area estimation in their areas, with support from methodology and geography staff as needed, is a natural choice since they should be most in touch with the data requirements and data limitations in their subject areas. More of an issue is how to package data for small areas for dissemination to users. Who should be responsible for pulling together data from different subject-matter areas for a particular small area? Should this be a regular program, or something that is done 'on demand'? Here there are different models to choose from – and Statistics Canada has tried most of them over the years.

At some periods in the past a division focussing on regional or urban statistics has existed to provide a regional focus for statistical data. At times, the census program, which is of course the richest source of small area data, has spearheaded the production of small area data profiles. At other times, an inter-divisional project has been used to manage a program of profiles for electoral districts or for other geographic areas. At the same time, regional office staff have played a key role in pulling together information for small areas in response to client requests. None of these arrangements has been ideal. The production of profiles has typically been a labour-intensive task requiring a broad subject-matter understanding and a lot of searching and manipulation of data. Despite the existence of standard geographic areas, the combination of data based on several different geographic bases is usually an issue. Ensuring that data for a large number of small areas are properly matched and collated can be an arduous quality assurance challenge.

Pre-planned profiles on paper were never overly successful. As a result, a strategy of maximizing responsiveness to client demands as they arose was preferred. With recent advances in technology, and broader coverage of small area data in the corporate database, a more automated approach is possible. A component of the Statistics Canada website (www.statcan.ca), called Community

Profiles, and largely based on Census of Population data, is our most recent attempt to make small area data more accessible and promises to be a precursor of future directions in this field. Some health data for health districts are already included, and certain other non-census sources of community data are under consideration.

8. CONCLUSIONS

The production of small area statistics by an NSO raises issues that are qualitatively different from those faced in its regular production of national, provincial or other large area data. The statistical theory that makes data based on sufficient individual measurements inherently reliable for large areas (ignoring bias for the moment) begins to break down for smaller areas. Unless a current census or administrative source with full coverage is available, this means that the NSO has to resort to some model-based help in order to provide estimates. Since alternative models can produce different estimates, a degree of arbitrariness is introduced into estimates, and this may be seen by some as undermining the objectivity of a NSO and its methods. The fundamental principle of openness and transparency about methods, including the choice of any models used and the impact of different assumptions, takes on even greater importance in the domain of small area estimation.

On top of this, an NSO should expect that small area estimates will come under more focused scrutiny than do many large area estimates. Though large area estimates receive broader attention, few individuals have the capacity to confirm or refute an estimate at the national level. But at the local level there will be many who think they know what is going on in their town. And typically small area estimation does not work uniformly well for all areas. The argument that a method works well on average will not quell criticism from those areas where it has not worked well – unless it has also worked to the local advantage! The NSO has to be prepared for the double jeopardy of weaker estimates under closer scrutiny.

If that is not enough already, confidentiality considerations loom larger at the small area level. The very fact that estimates are being produced for local areas highlights the potential for identification of individuals even though the NSO has taken sufficient precautions to prevent such disclosure. Some users of small area data for marketing purposes do not help the situation by implying in their advertising that they can target mail to households based on individual or household characteristics, when they are actually using small area data to distinguish neighbourhoods. Some methods of small area estimation require record linkage which may also raise privacy concerns. Again a policy of openness and careful review of all such applications, at a senior level and before they begin, is necessary to ensure that the public benefit outweighs any privacy invasion.

Despite these potential difficulties, the demand for small area data remains high, technology offers new approaches to the management and dissemination of small area data, and methodological work on small area estimation is an active research area among statisticians. While small area data will generally not be an NSO's first priority, the relevance of its statistical programs will be magnified many times if it is able to cater to the most important small area data needs.

ACKNOWLEDGEMENTS

This paper was originally prepared to introduce a discussion on small area statistics among Heads of National Statistical Offices at the June 2001 Conference of European Statisticians in Geneva. The author wishes to acknowledge the contributions of many individuals at Statistics Canada who provided input to this paper, as well as helpful suggestions from a referee.

REFERENCES

- ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Proceedings of the 52nd Session of the International Statistical Institute*. Helsinki.
- ALEXANDER, C.H. (2002). Still rolling: Leslie Kish's "Rolling samples" and The American Community Survey. *Survey Methodology*. 28, 1, 35-41.
- BAYARRI, M.J., and BERGER, J.O. (2000). *P* Values for composite null models. *Journal of the American Statistical Association*. 95, 452, 1127-1142.
- BRACKSTONE, G. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology*. 13, 1, 29-43.
- DURR, J.-M., and DUMAIS, J. (2002). Redesign of the French Census of Population. *Survey Methodology*. 28, 1, 43-49.
- FAY, R.E., and HERRIOTT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *Journal of American Statistical Association*. 74, 269-277.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on Statistical Disclosure Limitation Methodology (Statistical Policy Working Paper #22). Washington, D.C., Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.
- GAMBINO, J., and DICK, P. (2000). Small area estimation practice at Statistics Canada. *Statistics in Transition*. 4, 597-610.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*. 9, 55-93.
- ISNARD, M. (1999). *Alternatives to Traditional Census Taking: The French Experience*. Paris: INSEE.
- JABINE, T.B. (1993). Statistical disclosure limitation practices of united states statistical agencies. *Journal of Official Statistics*. 9, 2, 427-454.
- KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*. 16, 1, 63-71.
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*. 14, 31-46.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*. 35, 365-384.
- RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*. 25, 2, 175-186.
- REDFERN, P. (1989). European experience of using administrative data for censuses of population: the policy issues that must be addressed. *Survey Methodology*. 15, 1, 83-99.
- SCHAIBLE, W.L. (1996). (Ed.) *Indirect Estimators in U.S. Federal Programs, Lecture Notes in Statistics*. New York: Springer-Verlag, 108.
- SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*. 20, 3-14.
- ZAYATZ, L., STEEL, P. and ROWLAND, S. (2000). Disclosure limitation for Census 2000. *Proceedings of the American Statistical Association, Section on Government Statistics and Section on Social Statistics*. 67-71.

The Importance of a Quality Culture

DENNIS TREWIN¹

ABSTRACT

The reputation of a national statistical office (NSO) depends very much on the quality of the service it provides. Quality has to be a core value – providing a high quality service has to be the natural way of doing business. It has to be embedded in the culture of the NSO.

The paper will outline what is meant by a high quality statistical service. It will also explore those factors that are important to ensuring a quality culture in a NSO. In particular, it will outline the activities and experiences of the Australian Bureau of Statistics in maintaining a quality culture.

KEY WORDS: Continuous quality improvement; National Statistical Office.

1. INTRODUCTION

Fellegi (1996) provides a strong argument that the trust in the national statistical agency is how most users judge the quality of its statistical products.

“Credibility plays a basic role in determining the value to users of the special commodity called statistical information. Indeed, few users can validate directly the data released by statistical offices. They must rely on the reputation of the provider of the information. Since information that is not believed is useless, it follows that the intrinsic value and usability of information depends directly on the credibility of the statistical system. That credibility could be challenged at any time on two primary grounds: because the statistics are based on inappropriate methodology, or because the office is suspected of political biases.”

Trust will not happen unless the culture is right. Culture is a word with many meanings but I am interpreting culture as “the way we do things”. Core values are important to this. They cannot be just statements hanging on the wall. They have to be understood. They have to be reflected in behaviours, particularly by leaders of organizations.

The Australian Bureau of Statistics (ABS) places great reliance on adherence to its core values. More than anything, they distinguish us from other survey providers in Australia. The core values are:

- Relevance – regular contact with those with policy influence, good statistical planning, which requires a keen understanding of the current and future needs for statistics, are essential, as is the need for statistics to be timely and relatable to other statistics.
- Integrity – our data, analysis and interpretation should always be objective and we should publish

statistics from all collections. Our statistical system is open to scrutiny, based on sound statistical principles and practices.

- Access for all – our statistics are for the benefit of all Australians and we ensure that equal opportunity of access to statistics is enjoyed by all users.
- Professionalism – the integrity of our statistics is built on our professional and ethical standards. We exercise the highest professional standards in all aspects of ABS statistics.
- Trust of providers – we have a compact with respondents; they are encouraged to provide us with accurate information and we ensure that the confidentiality of the data provided is strictly protected. We keep the load and intrusion on respondents to a minimum, consistent with meeting justified statistical requirements.

Adherence to core values is just one element of maintaining a quality culture. Part 2 discusses the key steps the ABS uses to maintain a quality culture.

It is now widely recognized that quality is much more than accuracy (*e.g.*, Brackstone 1999 and Carson 2000). In Part 3, the different dimensions of quality are discussed before identifying in Part 4 what I think are some of the major quality challenges for the ABS over the medium term. Many of these will be shared by other national statistical organizations.

2. TOWARDS A HIGH QUALITY STATISTICAL SERVICE

Quality assurance is a responsibility of all staff in the ABS. There is no central “quality management” group although Methodology Division is encouraged to be our conscience on quality issues – a role it takes on with

¹ Dennis Trewin, Australian Statistician, Australian Bureau of Statistics.

enthusiasm, sometimes to the annoyance of others. However, that is a good sign – they are provoking debate on some of the more difficult quality issues. Support from senior management for this type of role is very important.

The key strategies for ensuring a high quality are described under six broad headings.

- A high degree of credibility for the ABS and its outputs.
- Maintaining the relevance of ABS outputs.
- Effective relationships with respondents.
- Processes that produce high quality outputs.
- Regular review and evaluation of statistical activities.
- Staff who are skilled and motivated to assure the quality of ABS outputs.

2.1 A High Degree of Credibility

Credibility is fundamental to the effective use of official statistics. Credibility arises from a system of statistics which provides an objective window upon the condition of a nation's economy and society.

The legislative framework within which the ABS operates is an important pre-condition for the integrity of Australia's official statistics. The Australian Statistician (*i.e.*, the chief executive of the ABS) is guaranteed considerable independence by law. This helps ensure that the ABS is, and is seen to be, impartial and free from political interference. In particular, the independence of the Statistician supports his objectivity in determining the statistical work program and determining what statistics are published. Although the legal authority is there, it still needs to be reflected in the way senior staff behave.

Government statisticians must not just apply professionalism skills to their work; they must also be seen to adhere to high ethical standards, especially with respect to objectivity and integrity. We are frank and open when describing our statistical methods to users; we publish information about our performance – for example, in terms of both sampling and non-sampling errors, and revision histories for key series; we are willing and able to identify and address user concerns regarding quality; we are receptive to objective criticism and prepared to respond quickly even if the problem is one of perception rather than reality. We promote good relationships with the media as they have a major influence on public opinion of the ABS and its outputs. Also, most Australians find out about official statistics through the media. We engage in other user education activities aimed at fostering intelligent use of official statistics.

The fact and perception of ABS objectivity are reinforced by our policies of pre-announcing publication dates for main economic indicators, allowing very limited pre-release of publications (the details of which are in the

public domain), and making special data services available on an even handed basis to all.

2.2 Maintaining the Relevance of ABS Outputs

There can be, of course, tension between (on the one hand) being responsive to changing policy needs and (on the other) maintaining the continuity of a system of statistics that can objectively monitor performance. Senior staff of the ABS devote a great deal of attention to maintaining personal contact with key users, to gather intelligence about policy issues and emerging areas of economic, social and environmental concern. This includes regular meetings with the most senior staff of the government agencies responsible for policy. The Directors of our State offices have similar arrangements with State officials. That intelligence feeds into strategic planning and the reviews of national statistical programs.

The ABS has a range of other means for communicating with the users of statistics, to ensure that our products are relevant to their needs. For example, advisory groups representing users and experts in various fields provide valuable guidance to our statistical activities.

There may also be some tensions or trade-offs between the different aspects of quality. The ABS positions itself at the higher accuracy end of the information market, to protect the valuable ABS "brand name". But if, for example, there is an urgent demand for data in a new field, some aspects of quality may be traded off in order to achieve timeliness and relevance. Nevertheless, there is a "bar" below which we will not go. Because it is probable that the new statistics will be used to inform significant decisions or debate, the ABS makes very clear statements about the accuracy of the data to help users understand how they can be used. On occasion, such new statistics may be differentiated from our other products by labelling them "experimental" or releasing as an information or occasional paper, rather than a standard publication. We regard this form of branding as very important to reliable interpretation of our statistics.

2.3 Effective Relationships with Respondents

An official statistical agency must maintain good relations with respondents, especially trust, if it wants them to co-operate and provide high quality data. The ABS approach includes – explaining the importance of the data to government policy, business decisions and public debate; a policy of thoroughly testing all forms before they are used in an actual survey; obtaining the support of key stakeholders; minimizing the load placed on respondents particularly by using administrative data where possible; and carefully protecting privacy and confidentiality.

The ABS monitors and manages the load it imposes on both households and businesses; we have developed 'respondent charters' for both groups. As well, a Statistical Clearing House has been set up within the ABS to

coordinate surveys of businesses across government agencies (including the ABS), to reduce duplication and to ensure that statistics of reasonable quality are produced.

All ABS forms and collection methods are tested to ensure that the data we seek are available at reasonable cost to respondents, and the best available methods are used to collect them. For business surveys, our units model, classifications and data items, are designed to be as consistent as possible with the way businesses operate. This now corresponds closely with their reporting for taxation purposes, making it easier to integrate survey data with data collected for taxation purposes. For household surveys, the extensive use of cognitive testing tools within the ABS, and the establishment of a questionnaire testing laboratory, have helped to improve quality and to reduce respondent load. Standards for form design and form evaluation are set out in manuals and are promoted and supported by experts in form design.

The ABS uses efficient survey designs to minimize sample sizes to achieve a specified level of accuracy, and hence total reporting load; we also control selection across collections to spread the load more equitably. To take advantage of current reforms of the Australian taxation system, the ABS is seeking every opportunity to improve the efficiency of our sample designs, through the use of taxation data as benchmarks, as well as using it as a substitute for some of the data now gathered through direct collections. We have changed the business unit structure used in our surveys to make it consistent with the structure used for taxation purposes.

For household surveys, the introduction of computer assisted interviewing has helped to streamline interviewing procedures, reduce respondent load, and improve the quality of data collected.

2.4 Processes that Produce High Quality Outputs

The quality of ABS statistics is underwritten by the application of good statistical methods during all stages of a collection including the design stages. The ABS has a relatively large Methodology Division (about 120 staff) which reports directly to the Australian Statistician. The Division is responsible for ensuring that sound and defensible methods are applied to all collections and compilations. The Methodological Advisory Committee, a group of academic experts, provides independent reviews of our statistical methods.

The ABS puts substantial effort into developing statistical standards, including concepts, data item definitions, classifications, and question modules. All ABS surveys must use these standards. The standards are supported by relevant data management facilities to ensure they are accessible and to make it easier to use standard rather than non-standard approaches.

Sample design and estimation methods are the responsibility of the Methodology Division. Where possible, a "total survey design" is used – accuracy requirements are set

according to the intended use of the data, and accuracy is measured in terms of both sampling and non-sampling errors. For example, in business surveys total survey design guides the allocation of resources to the intensive follow up of non-respondents or the editing of questionnaires; the effort for reducing non-sampling errors is optimized according to the impact of errors on overall quality. The cost to data providers is also taken into consideration. The "total survey design" has to be approved by a senior ABS committee before it is implemented.

In recent years, the ABS has made substantial progress by applying standardized best practice across surveys. For example, business surveys based on the business register now draw their frames at a common date each quarter, and use a common estimation method to ensure all collections have a consistent and complete coverage. Standard rules are adopted for frame maintenance, field collection and estimation, and generalized processing facilities are available to support the use of these rules. Standard methods are used to allow for "new businesses" not yet included on the survey frame. The ABS is thereby able to increase the coherence of estimates across different business surveys.

For household surveys, a master sample system has been adopted since the mid 1960's. The system is updated regularly after each five-yearly census, and has been the cornerstone for ensuring the accuracy of statistics collected from household surveys.

Achieving quality in surveys is easier when computer systems support current best practice. The ABS has invested in generalized tools. They have been developed for all major processing steps of both business and household surveys, including sample frame management, data input and editing, imputation, estimation and aggregation.

The ABS embraces a rigorous continuous quality improvement approach wherever appropriate. The Australian Population Census is a classic example of raising quality through a strategy of measuring quality and involving all staff in examining and devising solutions to quality problems. This approach was applied very effectively at the data processing centre for the 1996 and 2001 Censuses. In both cases, the centre achieved significant budget savings, better quality and an improvement in timeliness. Continuous quality improvement is also applied to the coding of businesses on the business register, and to many other ABS processes.

At the output end of collections, each subject group is required to confront its data with other ABS data and with external information, to ensure the coherence of our statistics. The key macroeconomic data have to be "signed off" by the national accountants in meetings established especially for the purpose of clearing the statistics. The national accountants then have an obligation to use this data without further adjustment in the compilation of the accounts, enhancing consistency between the national accounts and source data collections. More generally, confrontation of different data sources is undertaken by our

national accountants through use of an 'input-output approach' to compiling national accounts estimates. The new methodology has led to more consistent accounts. Furthermore, the data confrontation and balancing process at detailed levels have helped to identify data deficiencies. Information about quality is fed back to the economic collection groups and is resulting in a more focused approach to improvements in the quality of source data.

One important quality improvement initiative that the ABS has pursued is the development of an Information Warehouse to manage and store all of our publishable data. By drawing together different datasets into a single database, the Warehouse enables our statisticians to confront statistics produced from different collections. Furthermore, all forms of publication, be they paper based or electronic, are to be produced from a single data store, with the objective of ensuring that the same data released in different products, and at different times, are consistent.

Another important element of quality management is documentation. Good documentation supports review activity and facilitates the dissemination of quality information to users, so they can assess the fitness of the data for the purposes they have in mind. As part of the Information Warehouse initiative, the ABS can now enforce standards for documentation of the metadata that describe concepts, definitions, classifications and quality.

A relevant and responsive statistical service must do more than provide data to clients. The ABS has recently strengthened its analytical ability. A team of analysts has been set up to develop new measures of socioeconomic concepts, to explore relationships between variables and to prototype new analytical products. The expanded program of analysis work is expected to deliver significant benefits in the form of insights into data gaps and quality concerns.

2.5 Review and Evaluation of Statistical Activities

Each ABS area is responsible for continuous quality review and improvement. For statistical collection areas, quality management is supported by sets of performance indicators. A standard set of measures has been developed to permit a comparison of quality across collections. Tools are now being developed to calculate these measures as part of our normal survey processes, and the Information Warehouse will allow us to store and display the measures. The key indicators are also included in the annual reports each Branch makes to the ABS Executive for review.

Quality measures are of interest to the users of statistics. The Information Warehouse will improve users' access to information about quality issues. As well, the ABS places high priority on helping users understand the quality of data and their implications for them, and has adopted active education strategies to promote such understanding. As highlighted in Lee and Allen (2001), there is much to do to improve user understanding of quality.

Each ABS household survey now includes an evaluation program which reviews the effectiveness and efficiency of

all survey activities and assesses the extent to which the data are used by clients. The Statistical Clearing House conducts a review of each ABS business survey. These initiatives ensure that all collections are subjected to at least a basic evaluation, and brings to light opportunities for improvements to quality and efficiency.

As well as making internal comparisons of performance across its own collection areas, the ABS has established a benchmarking network with overseas statistical agencies; the aim of the network is to share information about survey design, processes and costs. The benchmarking exercise is providing very useful guidance to the ABS's efforts to improve its processes and outputs.

2.6 Skilled and Motivated Staff

The ABS could not provide high quality information to its user community if it did not employ people who bring skills and energy to our statistical work. The staff are responsible for implementing the strategies discussed above. They must take a professional approach and be committed to the development of new methods, to continuous quality improvement, and to the open discussion of methods and quality issues.

Quality improvement and on-going statistical work compete for the time and energies of our staff. The ABS approach is, as far as possible, to integrate quality work with on-going processes and systems. We emphasize to staff that quality management is a corporate priority and ensure that tools and resources are made available to support it. In particular, the ABS is implementing a tighter approach to project management; this is being supported by manuals, systems and training.

Statistical training plays an important role in maintaining and improving quality. The ABS is always searching for new, more effective, approaches to skills development. An important element of our performance management system is a focus on identifying and addressing individuals' development needs.

Relationships with other national and statistical agencies are a very important element of the ABS efforts to improving official statistics. The ABS is committed to using international standards; we take advantage of the wide range of expertise embodied in those standards. On the other hand, there is an obligation for us to make a positive contribution to the development of the standards. In doing so, we try to take account of the interests of the Asia/Pacific region as well as those of Australia. With ever increasing globalization of economic activity and the pursuit of world wide social goals, the compatibility between Australian statistics and those of other countries, is an important element of quality. The ABS maintains strong links with many overseas agencies. We are fortunate that there is a lot in common in the challenges we face and there are great benefits from sharing experiences with other statistical agencies.

3. DIMENSIONS OF QUALITY

Figure 1 is taken from Lee and Allen (2001). Among other things, it neatly summarizes, on the left hand side, three existing frameworks for judging quality. There are some differences with the descriptors used but basically they are providing the same message – there is much more to quality than accuracy. This is now widely accepted although it was not so long ago that discussion of the quality of a statistic focussed on its accuracy and the sampling variability in particular.

There are several messages in the right hand side of Figure 1.

- (i) There are many different ways of compiling official statistics – from modelled data/analytical outputs to censuses and sample surveys. In Australia we are making greater use of administrative data, systems of accounts (linked to the national accounts) and model based and other analytical methods to produce statistical outputs, compared with five years ago. The quality challenges differ between the different means of compiling statistics.
- (ii) There are several groups of activities associated with statistical outputs – from “frameworks, concepts, standards and classifications” through to “services/dissemination”. Each is important in its own right and has its own quality challenge.

- (iii) The performance of a National Statistical Office is extremely important to its quality image as recognized in the opening quote of the paper. A number of the elements are specified in Figure 1. All are important. Indeed you cannot have a high performing statistical office unless you rate well against each of these elements; including management and financial performance.
- (iv) There are other elements such as institutional settings (*e.g.*, legislation) which are also important.

The main purpose in describing the above is to emphasize that the list of quality challenges for a national statistical office is very large. All have to be tackled in some way – this would not be possible unless you have a quality culture, *i.e.*, attention to quality is the responsibility of all staff. There are many “moments of truth” to genuinely test whether a quality culture exists or not.

4. CURRENT QUALITY CHALLENGES AT ABS

Psychologists say that it is difficult to grasp more than seven points at one time so the remainder of the paper is limited to identifying seven major quality challenges for the ABS.

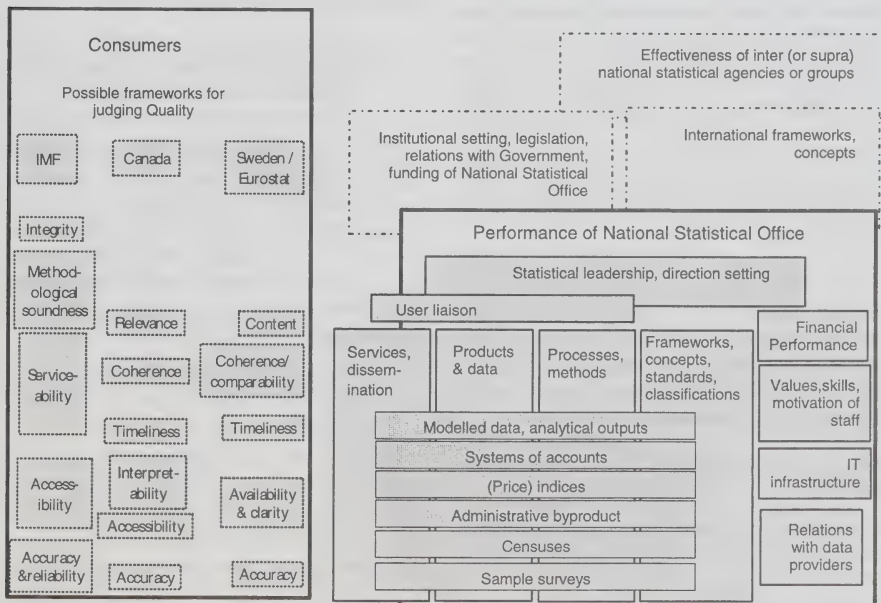


Figure 1. A Framework for Assessing Quality

- (i) The increasing use of large, but imperfect, administrative and transactional data bases for compiling official statistics.
- (ii) Increasing user expectations raising the quality “bar”.
- (iii) Managing the tension between improving business processes (which can mean removing those responsible for statistical outputs from direct involvement with input processes) and maintaining or improving the quality of statistical outputs.
- (iv) Quality assurance on electronic outputs.
- (v) The presentation of statistics on the internet, including the need to educate the user community on quality of official statistics.
- (vi) Managing the transfer of knowledge and skills with an ageing senior management team, many of whom will retire over the next 5 years.
- (vii) Use of international statistical standards to maintain comparability where the standard may not be the most appropriate for national statistics.

4.1 Increasing Use of Administrative/Transactional Data Bases

We have used administrative data bases for many years (*e.g.*, vital registrations for births and deaths, customs for trade data) to compile official statistics. Others have been used to develop frameworks for statistical collections. The issues at hand are the increasing availability of these data bases, their under-utilization for statistical purposes, and taking advantage of the potential to link across data bases and ABS collected data sets using a common identifier (*e.g.*, the Australian Business Number for business statistics).

Examples of administrative data bases that are becoming available are extended personal and business income tax data bases, health insurance transactions, and details of those on income support.

Transactional data bases are becoming available, although not in readily accessible form. Data bases of particular interest to the ABS are scanner data bases from retail outlets and eftpos (*i.e.*, electronic fund transfers between customers and retailers) data bases.

There are some particular advantages in using administrative or transactional data bases:

- they reduce the compliance cost we impose on respondents
- they are often “censuses” and therefore provide scope for producing detailed data sets (*e.g.*, by geography)
- they often have a longitudinal element (*e.g.*, tax data) to support this form of analysis

- they often contain an identifier which facilitates analysis across data sets (*e.g.*, the *Australian Business Number* will facilitate analysis across business tax data sets, customs data, and ABS surveys)
- they might be cheaper than directly collected data sets.

There are negatives of course – for example, the definitions may not be consistent with the preferred statistical concepts; less attention may have been given to incoming quality; and they may be out of date. Managing privacy aspects is a particularly important element. Although our motives are entirely honourable, and are in the public interest, matching data bases is a sensitive issue and ignored at our peril. Many of our users, particularly those in the academic community, are not as sensitive to these concerns.

There is also the question of whether the ABS should produce the statistical outputs or the agency responsible for the data sets. A number of issues come into consideration – the importance of the outputs to the national statistical service, costs, the extent to which quality can be managed and the basic question of whether the administrative agency is prepared to give up custodianship. Only the most important data sets will be brought into the ABS for compiling official statistics; for the others, we will work with the administrative agency to help them deliver “fit for purpose” statistical outputs into the public domain.

What have been our key responses to this important quality management issue?

- We are developing protocols for the publication and management of data from administrative sources. Associated with this is the promotion and support of good statistical and data management practices.
- For each statistical field, we are preparing information development plans in conjunction with other stakeholders which identify those areas of greatest importance and set out specific activities which will lead to increased availability of non-ABS data, particularly quality management issues.
- We are actively promoting good practice in information management.
- A major investment project has been the greater utilization of taxation data to provide cost-effective statistics.
- We are investigating methods for assuring the quality of the very large but imperfect data sets that are available through administrative and transactional data holdings.

4.2 Increasing User Expectations

User expectations on quality are changing – they are much higher than what they were as recently as 5-10 years ago. This trend is likely to continue. The increasing

globalization of financial markets will mean that key macroeconomic statistics have international, as well as national prominence.

There is a perception that statistics have become more volatile. In some cases they have because the underlying phenomenon has become more volatile. However, we do not believe statistical measurement methods are a significant contributing factor – in most cases methodological developments have led to improvements although the perception may be different. For example, the volatility in the key national accounts series is considerably less than what it was 10–15 years ago yet this is quite different to the perception of some users.

We also receive more criticism of inaccuracies in very detailed data (*e.g.*, Population Census tables) than previously. Again, it is not that the quality is deteriorating – it is that the expectation is higher.

We have to accept that “the bar is rising” and do what we can to improve quality to the expected level. That is not always possible of course so managing expectations is important. This can be done by:

- providing good explanations of the strengths and weaknesses of particular data sets;
- talking to key users whenever possible about the strengths and weaknesses of data series;
- responding to their informed criticism (seek partnerships in improving quality *e.g.*, in our detailed foreign trade statistics we openly seek feedback from users on the quality of the statistics); and
- providing as much explanation as possible for statistics that might seem unusual or different to expectation.

4.3 Improving Business Processes

Like several statistical organizations, the ABS is looking at how it might use new technologies, and other elements such as increased access to taxation data, to improve the efficiency of its business statistics processes.

We are also investigating the business processes associated with household surveys, particularly as increased use is made of computer assisted interviewing (CAI). However, in this section the paper will concentrate on the changes we are making to the way we manage business statistics to describe this particular quality challenge.

A team was set up to look at the possibilities. As a consequence, a number of significant changes were agreed to – this is to be known as the Business Statistics Innovation Program. We are looking at revised business processes that will be in place for at least 10 years and will yield a significant return on the investments required to set it up. We will:

- extend the responsibilities of the Business Register Unit to capture and store taxation data with a direct link to the Business Register through the Australian

Business Number (ABN). The ABN is now allocated through the taxation registration scheme and is available with most business transaction data bases. The data will be stored in a way that it can be used by the various ABS statistical areas to compile statistics directly from taxation data or in combination with ABS survey data;

- improve the way we manage business respondents – this will include some preference in how they provide data to us;
- set up an input data warehouse, with the Australian Business Number as the link across the various data sets;
- establish a business statistics processing environment based around the input data warehouse; and
- increase centralization of a number of the functions associated with compiling business statistics.

We can see the positives in these developments – more efficient delivery of business statistics, enhanced use of taxation data and other administrative data, data bases that support a wider range of statistical analysis. However, it will reduce the level of contact that statistical output areas have with their input data sources. What impact will that have on quality? What strategies can we deploy to mitigate the impact? These are important questions that we will have to answer. It is the main risk we will have to manage in implementing the Business Statistics Innovation Program.

4.4 Quality Assurance on Electronic Outputs

Great care is taken on the quality of our paper products. This has been built on many years of experience. Our record is good and the quality assurance processes well embedded in the way we go about our business. Yet, more and more of our user community receive their data in electronic form only. They will make analyses based on these outputs often leading to important decisions being made. It is just as embarrassing to us to have errors in electronic outputs as to have them in paper outputs.

Our quality assurance procedures for electronic outputs are not as sophisticated, but they are evolving. The key responses have been as follows:

- Our data warehouse supports the storage of all the objects associated with the dissemination with a particular set of statistics, including data cubes and meta data.
- Statistical areas are asked to approve each object – they are individually developing their own techniques for quality assurance (but sharing ideas on best practice).
- A publishing system has been developed to support the simultaneous release of all outputs. If they are delivered from the same set of objects, there is less chance of inconsistency between the outputs.

4.5 The Presentation of Statistics on the Internet

Ultimately the user can only make judgements about the fitness of a statistical output for their purposes. These vary of course and what might be fit for one purpose may not be for another. There is an obligation on us to provide a range of supporting information on data outputs, including that on quality, so that the statistical users can make their own judgements on fitness of use. There are a number of existing, well proven practices relating to declarations about the quality of statistics. These activities are now a routine part of existing dissemination practices. They include:

- Concepts, Sources and Methods publications that describe in detail the methods used to compile major statistical outputs. These are available on our web site as well as on other media.
- An assortment of Information and Working Papers, and feature articles in publications, which are used to draw attention to issues specific to particular outputs or changes that are being made to their compilation methods.
- A policy of “no surprises” when there are significant changes to the methods used for the compilation of statistical series. As well as Information Papers *etc*, if there are important changes to statistical series, we embark on a program of seminars and bilateral discussions with key users to explain the changes and the reasons for their changes.
- Material on methods is included in all our publications. The ordering and physical presentation of this information is according to agreed standards. These were developed following research undertaken for us by a communications consultant on how our users use the material in statistical publications.
- The analysis section of our publications includes material that explains, among other things, large or unusual movements in our statistical series. Often this will be based on information that is only available to ABS staff through their contact with respondents or their intimate knowledge of the methods used in compiling statistics. Our User Groups have advised that this is one of the most valuable forms of analysis that we can undertake.

We believe that our key users have a reasonable understanding of the quality of the statistics they use. However the increased reliance on electronic dissemination poses new challenges. In one sense this move provides a wonderful opportunity to present a range of information on quality that is easily accessible through a few well-designed “clicks”. But because information about the quality of the statistics is “not in your face” like it can be in hard copy publications it is easier for users to avoid the key messages

that you are trying to convey. The real challenge for us is to develop methods for presenting quality in a way that is not easy for users to avoid the main messages we want to convey.

One means of doing this may be to provide separate messages that draw attention to particular information you want to transmit on quality. These could be automatically activated as particular statistical series are accessed or could be delivered by a separate email message. Research is required into the most effective means.

Lee and Allen (2001) have described some of our research work to date on this issue. The work is still at the exploratory stage. Things that are being investigated are:

- Usability testing of how users prefer to access information on quality.
- Showing leadership and developing user education programs on how to use information on quality. A trial version of the is now available.
- The development of four prototype tools to assist users understand the quality of particular statistics. The four prototype tools are “Quality Issue Summaries”, “Quality Measures”, “Data Accuracy” and “Integrated Access to Data and Metadata”.

More details are available in Allen (2001).

4.6 Managing the Transfer of Knowledge and Skills

Like several other national statistical organizations, many of the ABS management team, and other senior staff, are aged in their 50’s. Some have retired in recent years. Others are expected to over the next few years. If managed correctly, this is a great opportunity to refresh the organization through providing new blood to management positions. These will normally be younger staff who will bring new ideas and energy into the management team.

On the other hand, experience and know-how will be lost. Both sides of this equation need to be managed carefully. Our strategy is as follows.

- We have developed special programs for those staff with potential. Specifically, they undertake a leadership and management development program which has been specially customized for the ABS. Staff are chosen for these programs by senior managers. You cannot select yourself to be a participant in the program. Furthermore, after staff have completed the program they can be expected to be chosen for a special assignment or rotated to a new position. The underlying philosophy is that the best way of learning is to obtain a variety of work experiences. A very high proportion of recent promotions to senior management positions have been participants in these programs. So far this has helped us to adequately cover the gaps created by a larger number of retirements than in the past.

- We retain links with retired ABS staff through a variety of informal and formal means (*e.g.*, social functions, including them on the distribution list for ABS News, *etc.*). Their knowledge is accessible if required.
- We have placed a stronger emphasis on knowledge management, using the facilities of our groupware product (Lotus Notes), means that key parts of our work are well documented and easily accessible.
- We have made substantial moves to standardize methods and systems meaning there is less dependence on local knowledge.
- For some key positions (*e.g.*, Director of National Accounts) we ensure shadowing of work prior to the retirement of the incumbent.

To date we have managed this transition well. We have been able to adequately fill vacant senior positions and at the same time refresh the organization by promoting staff with fresh ideas. There is a need to remain adroit.

4.7 Use of International Standards

Our starting position is that where international standards exist we should use them. This has not always been the case. For example, although our industrial classification has been loosely based on ISIC, and a concordance developed with ISIC, the classification is largely homegrown reflecting the specific interests of Australia and New Zealand. We have agreed to use the 2007 version of ISIC, at least for the upper two levels, with variations at lower levels only where there are specific circumstances that justify it.

There are often pressures on us to divert from international standards. Sometimes this is to make the Australian situation look better. In other cases, such as with the ILO unemployment definition, the pressure is because the international definition does not seem to reflect the real situation in Australian circumstances. We resist these pressures but it is important that we have a well documented international standard as a reference point to justify our position. Nevertheless, where diversions from the international standard are made on an exception basis, they need to be well documented with a clear explanation of the reason. In cases where there is a need to have information on a basis other than the international standard our position is that we should publish statistics on both bases. The headline figure would still reflect international standard as increasingly the Australian situation is being compared with that of other countries and it is important that it is done on a comparable basis. For example, this approach is being taken to satisfy the demand for underemployment data and to reduce criticisms of the ILO unemployment definition.

There is a tension that needs to be managed but if we are serious about the importance of international comparisons it is imperative that international standard is the main

guiding light in developing the concepts, sources and methods used in Australia. For these reasons we regard it as a priority to make a significant contribution to the development and revision of international standards.

5. CONCLUSION

We would all agree that attention to quality is a fundamental aspect of our operation. In this paper, we have attempted to show that there are many dimensions to quality. This same message is clear from the frameworks for quality that have been developed by other organizations, such as the IMF, Statistics Canada and Statistics Sweden. The consequence is that a quality organization depends on the actions of all its staff as all can have an impact on quality in one way or another. It cannot be left to a work group with designated responsibility for quality. Therefore, quality can only happen if there is a genuine quality culture within the organization. The paper attempts to describe how we achieve this within the ABS. Nevertheless, it is important to have someone who performs the role of the corporate conscience on quality. We have given this responsibility to the Methodology Division and made the Chief part of the ABS Executive team so that it is easier for key messages to be conveyed to the senior managers. Among other things they draw attention to the most important risks to quality or behaviours they see as contrary to our corporate objectives.

ACKNOWLEDGEMENT

Section 2 was partly based on a document produced by Frank Yu (ABS) for the 9th Conference of South Asian Statistician Offices.

REFERENCES

- ALLEN, B. (2001). Qualifying Quality – Issues of Presentation and Education. *Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada.
- BRACKSTONE, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*. 25, 2, 129-149.
- CARSON, C. (2000). Towards a framework for assessing data quality. The Proceedings of the Statistical Quality Seminar, Jeju Korea, Korean National Statistical Office and International Monetary Fund.
- FELLEGI, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*. 64, 165-197.
- LEE, G., and ALLEN, B. (2001). Educated Use of Information about Quality. *Bulletin of the International Statistical*.

Model Explicit Item Imputation for Demographic Categories

YVES THIBAUDEAU¹

ABSTRACT

We propose an item imputation method for categorical data based on a MLE derived from a conditional probability model (Besag 1974). We also define a measure for the item non-response error that is useful to evaluate the bias relative to other imputation methods. To compute this measure, we use Bayesian iterative proportional fitting (Gelman and Rubin 1991; Schafer 1997). We implement our imputation method for the 1998 dress rehearsal of Census 2000 in Sacramento, and we use the error measure to compare item imputations between our method and a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) at aggregate levels. Our results suggest that our method gives additional protection against imputation biases caused by heterogeneities between domains of study, relative to the hot-deck.

KEY WORDS: Nearest Neighbor; Conditional probability approach; Bayesian iterative proportional fitting.

1. INTRODUCTION AND BACKGROUND

Let S represent a demographic categorical count requested from a census, or needed to compute a survey statistic, and suppose S can be computed from the records of a survey file f , when the records are complete. Also, suppose f is ordered in such a way that proximity in the order of f corresponds to geographical proximity. Consider the situation where f includes records with unreported items. We propose to estimate S with $d(A(f))$, where $A(f)$ is an imputation method that produces a complete survey file, and $d(\cdot)$ estimates S by replacing the unreported items with their values imputed with $A(f)$. $A(f)$ is based on a likelihood that models transitions between two neighbors in f , and associations between the items to be imputed and the relevant domains of study (Cochran 1977, page 34) defined by partitions of the population. $A(f)$ is meant as an advantageous alternative to the popular sequential hot-deck (Kovar and Whitridge 1995), which is a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) that attempts to minimize geographical distance between a unit with unreported items and a suitable imputation donor, while also guaranteeing the distributional homogeneity of the observed and the imputed items with respect to each domain of study. When the domains of a same partition tend not to geographically overlap, borrowing imputation items from a near-by neighbor preserves homogeneity. But, when small domains tend to be dispersed within large domains, the methodologist faces a dilemma. Then, she must choose between hot-deck rules that lead to borrowing the imputed items from geographically close units, leaving the possibility of imputation biases reflecting the local heterogeneity between domains, and domain-specific rules, which guarantee distributional homogeneity by domain, but may not minimize geographical distance. $A(f)$ is an alternative designed to

preserve domain integrity, while also simulating the distributional profile of an imputation donor sharing some characteristics with a geographical neighbor. We motivate the design of $A(f)$ with examples and a theoretical description. In this section we review a classification of current hot-deck methods for item imputation with their operating principles, so that we can properly compare them with $A(f)$ in later sections. We also give details on the dress rehearsal of Census 2000 in Sacramento, our test bed throughout the paper.

Fay (1999), and Sande (1981) identify the sequential hot-deck (SHD) as the first category of hot-decks, which we call the "pure" SHD. They add a second category, the fixed-cell hot-deck (FCHD), which we call the pure FCHD. Fay defines a third category of hot-decks: the nearest neighbor hot-deck (NNHD). Chen and Shao (1997, 2000) give an abstract definition of the NNHD in terms of a measure of proximity $|\cdot|$, based on a covariate x . With the NNHD, a "donor" is any unit such that $|x_r - x_d|$ is minimal, where x_r corresponds to the receiving unit (receiver), and x_d corresponds to the provider of the imputations (donor). By constructing the appropriate measure, and defining a suitable x , we recover both the pure SHD and the pure FCHD as special cases of the NNHD. The pure SHD imputes a receiver item by replacing it with the corresponding item from the closest unit for which it was reported, in the order of f . The pure FCHD relies only on the value of variables that we call the class variables to divide the units between post-strata that are homogenous with respect to the items to be imputed. A donor is chosen at random from the same post-stratum as that of the receiver, irrespective of the order of f .

Fay (1999), and Fay and Town (1998) propose the concept of exchangeability to validate the NNHD. For categorical data two units in f are exchangeable if they are uncorrelated and identically distributed, given the

¹ Yves Thibaudeau, Mathematical Statistician, Statistical Research Division, US Census Bureau, 4700 Silver Hill Road, Stop Code 9100, Washington, DC 20233-9100. E-mail: yves.thibaudeau@census.gov.

information available prior to imputing. The operational assumption of the NNHD is that a unit and its nearest neighbor(s) are exchangeable. For the pure SHD it means two contiguous units in f are exchangeable. For the pure FCHD it means that units sharing the same values for their class variables anywhere in f are exchangeable. We define a third instance of the NNHD, which we call the hybrid sequential hot deck (HSHD). To guarantee exchangeability the HSHD requires proximity both in terms of the order of f , and in terms of the class variables.

We use the term "nearest neighbor" in the abstract sense of the NNHD, unless specified otherwise. We use the terms "closest neighbor" to designate the nearest neighbor of the pure SHD, and "closest complete neighbor" to mean the survey unit with no unreported items that is closest in the order of f . In the case of the Sacramento dress rehearsal, the Census Bureau uses a HSHD to estimate householder counts by tenure, race, origin (Hispanic origin), and sex. The householder, usually an adult, is unique for each housing unit, and is determined by the ages, relationships, and order of the persons on the census questionnaire. The HSHD substitutes unreported items with the values of these items corresponding to the last householder who reported them and is in the same post-stratum (Treat 1994). The sorted order of f maintains the proximity of geographical neighbors. The intent behind the HSHD is to define nearest neighbors who are close, both in geography and "in kind". Throughout the paper, we continue to use the term householder, although its meaning may extend to a generic survey unit.

The design of the HSHD is well suited for item imputation in populations geographically clustered by domain. Then the need for class variables is limited. But difficulties arise when the geographical boundaries between the domains begin to blur. Designing a HSHD with good discrimination power in those conditions is an attempt at walking a fine line between specifying enough class variables to account for heterogeneities between domains, and specifying too many, which could yield post-strata so narrowly defined in terms of domain that they don't capture the local geographical character of the receivers. Complicating the situation is the fact that the demographic composition of the population may change as the geography changes, and thus a particular scheme for the HSHD might need to be revised, as the geography changes. In the face of these difficulties $A(f)$ is innovative in the sense that, instead of searching for an ideal nearest neighbor, it generates imputations through a model-based simulation that integrates information relating to the local geography, as well as to domain partitions. $A(f)$ integrates both kind of information by calibrating the parameters of a log-linear model on the basis of the strength of the correlations between the covariates and the variables subject to imputation. Our parameter estimation strategy is the same as that of Zanutto and Zaslavsky (1995a, b). However, because they have access to a representative sample of complete non-respondents, these authors can obtain estimates of the

imputation probabilities by implementing a one-step EM algorithm (Dempster, Laird and Rubin 1977). In our situation, we don't assume access to a representative sample, and we implement the full EM algorithm. Implicitly we make an assumption of items "missing at random" (MAR) (Little and Rubin 1987, page 16).

To analyze the results obtained with $A(f)$, and to compare them with those of the HSHD, we derive error measures related to $A(f)$ based on approximations computed using a Bayesian algorithm first introduced by Gelman and Rubin (1991). There are fundamental objections to Bayesian methodologies. Fay (1992) shows that variance estimation based on multiple imputations (Rubin 1996) can lead to inflated estimates of variance, whereas in the same situation the jackknife estimator (Rao and Shao 1992) avoids biases. Meng (1994) suggests that Fay's example stems from a poor communication between an imputer who has specific model information, and an analyst who only has knowledge of the estimation process. In the language of Meng, this situation is uncongenial. While requirements for coordination between imputer and analyst are restrictive, imputation based on exchangeability also has dangerous pitfalls, as we show in section 2. In addition the Bayesian approach allows for asymptotic approximations of error measures through mechanical algorithms, while a strict frequentist approach might require tedious expansions, as we show in section 5.

Our objective is to present $A(f)$, and to show its comparative advantages over the HSHD, using the Sacramento dress rehearsal as an example. In this case f contains records for the 138,271 physically enumerated householders (Kostanich 1999), of whom 90,156 returned a census questionnaire by mail or were visited by an enumerator at a first attempt, and 48,115 were selected in a sample. We implement our method at the level of the tract, a connected unit of geography containing on average 1,300 householders in f .

The paper is organized as follows. In section 2 we illustrate the difficulties of designing a HSHD methodology that guarantees exchangeability. In section 3, we define $A(f)$, and in section 4 we present a likelihood for the model parameters. In section 5, we show how to implement $A(f)$ and derive a measure of error to make comparisons with the HSHD. Section 6 presents and motivates the basic model for the dress rehearsal, and section 7 gives results for both $A(f)$ and the HSHD in this case. In section 8, we summarize the differences and we make recommendations.

2. ASSESSING EXCHANGEABILITY WITH RESPECT TO A PARTITION BY DOMAINS OF STUDY

We illustrate the difficulties inherent in designing a HSHD that preserves exchangeability between domains of study (Cochran 1977, page 34) with an example, where tenure (ownership) is the measurement of interest, and the

relevant domains of study are defined by race. To impute tenure, the Census Bureau uses the class variable "household type" to post-stratify f in five post-strata defined by the presence/absence of a live-in spouse for the householder, and the size of the household (1, 2, 3+) (Wilson 1998). The intent is to define post-strata that establish distributional homogeneity in terms of ownership at the level of the post-stratum, rendering the domain boundaries of a relevant partition uninformative within each post-stratum.

We examine the post-stratum comprising all the householders without a live-in spouse, and living in households of 3 or more. We call it post-stratum 3. For the purpose of this example, we have removed from f all the householders with unreported tenure, and each nearest neighbor is exclusive to a single householder. Table 1 gives householder frequencies for eight exhaustive race-tenure categories for post-stratum 3. Table 1 also gives the rate of ownership for their nearest neighbors, cross-classified by their race and by the same eight race-tenure categories of the corresponding householders. We observe that, on average, when a householder is either in the Black-owner or in the Black-renter category, his nearest neighbor is at least 25% more likely to be an owner when this nearest neighbor is White, than when he is Black. It is tempting to explain this differential rate by geographical differences. However, table 2, which shows the rates of ownership of the householders in post-stratum 3, cross-classified by their own race and that of their nearest neighbors, reveals that in fact Blacks with White nearest neighbors have a slightly lower rate of ownership than Blacks with Black nearest neighbors. What this means is that, if the probability of not reporting tenure is constant for all Blacks, then imputing their tenure by

substituting the tenure of their nearest neighbor overestimate ownership for Blacks in post-stratum 3.

These distributional disparities between householders and their nearest neighbors reflect a lack of exchangeability. A McNemar test leads to a formal rejection of the exchangeability hypothesis. There are 1,784 Black householders with White nearest neighbors. In 1,187 instances, tenure is tied. Among the 597 non-tied cases, the owner is White in 396 cases. Under the exchangeability hypothesis, ownership goes to either race with probability one-half. But the proportion of Whites among the owners is eight standard deviations above one-half. This example illustrates the difficulties in designing a valid NNHD that maintains exchangeability. In the next section we present our imputation method, which is devised for this type of situation.

3. AN IMPUTATION METHOD BASED ON DEMOGRAPHIC TRANSITION PROBABILITIES

Besag (1974) describes the conditional probability approach to spatial processes. This approach gives a framework for probabilistically modeling the values of "sites", in terms of the values of their "neighbours" to construct a spatial process. Besag (1974) also suggests making a unilateral approximation to simplify this construction. Then, the value of each site depends only on a finite number of "predecessors". This approach is natural in our situation since f provides a unilateral ordering of householders who play the roles of sites and predecessors, in turn. Specifically, we construct a first-order process where each householder is a site, and the complete closest neighbor is

Table 1

Number of Householders and Rates of Ownership of the Nearest Neighbors in Post-Stratum 3 by Race of the Nearest Neighbor and Joint Race and Tenure of the Householder

	Race-Tenure Category of the Householder							
	White Owner	White Renter	Black Owner	Black Renter	Asian Owner	Asian Renter	Other Owner	Other Renter
Number of Householders in Post-Stratum 3	3,347	5,197	1,319	3,630	872	1,196	681	1,637
Rate of Ownership of the White Nearest Neighbors	0.556	0.564	0.562	0.299	0.561	0.287	0.540	0.163
Rate of Ownership of the Black Nearest Neighbors	0.379	0.189	0.427	0.211	0.443	0.202	0.471	0.158
Rate of Ownership of the Asian Nearest Neighbors	0.589	0.332	0.667	0.320	0.668	0.262	0.535	0.302
Rate of Ownerships of the Other Nearest Neighbors	0.423	0.251	0.497	0.237	0.595	0.177	0.463	0.152

Table 2

Rates of Ownership of the Householders in Post-Stratum 3 by Race of the Householder and Race of the Nearest Neighbor

	Race of the Nearest Neighbor			
	White	Black	Asian	Other
Rate of Ownership of the White Householders	0.415	0.358	0.384	0.337
Rate of Ownership of the Black Householders	0.257	0.264	0.304	0.267
Rate of Ownership of the Asian Householders	0.441	0.441	0.400	0.360
Rate of Ownership of the Other Householders	0.309	0.297	0.337	0.234

its only predecessor. In this set-up, the value of a site is the state of a householder, which we define shortly. We refer to the conditional probability for the value of a site given that of its predecessor as the transition probability from the state of the closest complete neighbor to the state of the householder. Our imputation methodology is based on the MLE of the transition probabilities at the level of a tract. In this section we describe the imputation methodology, and in the next section we introduce a likelihood for the transition probabilities.

Consider a population of householders in f representing a tract. Let Ψ represent a set of C categorical variables that characterize each householder. The variables are labeled $1, \dots, C$, and have respectively K_1, \dots, K_C categories. Let Ψ^\times denote the Cartesian product of the categorical variables in Ψ . Then, Ψ^\times is the state space of the householder and has K states, where $K = \prod_{i \in \Psi} K_i$. Similarly, let Ξ be the set of E categorical variables defining the closest complete neighbor in f . The variables are labeled $1, \dots, E$, and have F_1, \dots, F_E categories. Ξ^\times is the state space of the closest complete neighbor and has F states, where $F = \prod_{i \in \Xi} F_i$. The items represented in Ξ are also represented in Ψ . Let the state of the householder be $s \in \Psi^\times$, where s is a vector whose components represent the variables in Ψ . Similarly, $t \in \Xi^\times$ is the state of the closest complete neighbor. Under the assumptions above, let $P(s|t)$ represents the transition probability from t to s in the order of f . Now suppose a householder only reported the categorical variables in a subset $Z \subset \Psi$. Let $v \in Z^\times$ be the vector of reported variables. Let $\sigma(\Psi, Z, v) \subset \Psi^\times$ be the subset containing all the values of s , such that s agrees with v on the variables in Z . Define

$$P(s|t, Z, v) = \frac{P(s|t)}{\sum_{u \in \sigma(\Psi, Z, v)} P(u|t)}; \quad s \in \sigma(\Psi, Z, v). \quad (1)$$

To impute the items in the set difference $\Psi - Z$ according to $A(f)$, we roll dice weighted by the values of the MLE of $P(s|t, Z, v)$, for each householder in marginal state v and with closest complete neighbor in state t . Under our assumptions, the MLE of $P(s|t, Z, v)$ contains all the information available from f on the unreported items. In the next section we formulate a likelihood for $P(s|t, Z, v)$.

4. A LIKELIHOOD FOR THE TRANSITION PROBABILITIES

Let $N(t, Z, v)$ be the number of householders who only reported the items defining the marginal state v involving only the items in $Z \subset \Psi$, and with closest complete neighbor in state t . Let N be a vector with the $N(t, Z, v)$'s as its components, at the level of a tract. Let $P = [P(s|t)]$ be the vector comprising the $P(s|t)$'s ordered lexicographically by t and s . Based on the assumptions described above, we have the following likelihood for the transition probabilities.

$$L(N; P) = \prod_{t \in \Xi^\times} \prod_{Z \subset \Psi} \prod_{v \in Z^\times} \left(\sum_{s \in \sigma(\Psi, Z, v)} P(s|t) \right)^{N(t, Z, v)}; \quad P \in \Theta_P. \quad (2)$$

The running indices in (2) are t, Z, v , and s . If every item is reported, then Ψ is the only instance of Z with $N(t, Z, v) \neq 0$, for some t and v . In that case (2) is analogous to the likelihood of the transition probabilities of a first-order Markov chain (Bishop, Fienberg and Holland 1975 page 263). In general, we model Θ_P as a log-linear subspace. For this purpose it is more convenient to work with an expression equivalent to (2) that has a simpler algebraic representation. We introduce the nuisance parameter $U = [U(t)]$, where U is a probability vector, that is $\sum_{t \in \Xi^\times} U(t) = 1$, and $0 < U(t) < 1$, for all $t \in \Xi^\times$. U represents the prevalences of the states of the closest complete neighbors. Let $Q(s, t) = U(t) \times P(s|t)$, and $Q = [Q(s, t)]$. Then Q is a probability vector with $K \times F$ components lexicographically ordered by t and s . We set up Θ , the parameter space of Q , as a hierarchical log-linear model (Agresti 1990, page 143; Bishop, Fienberg and Holland 1975, page 67). Then, if we design Θ so that it includes the interactions of all orders between the variables in Ξ , (2) is equivalent to the following likelihood in terms of Q .

$$L^*(N; Q) = \prod_{t \in \Xi^\times} \prod_{Z \subset \Psi} \prod_{v \in Z^\times} \left(\sum_{s \in \sigma(\Psi, Z, v)} Q(s, t) \right)^{N(t, Z, v)}; \quad Q \in \Theta. \quad (3)$$

That is, if Θ has the architecture described above, a specific choice for Θ unambiguously defines Θ_P in (2), and since the items of the closest complete neighbor are always reported, the factorization $L(N; P) = L^*(N; Q) \times R(N; U)$ holds, for some $R(\cdot)$. (3) is easier to manipulate than (2) since it corresponds to the likelihood of the cell probabilities associated with a partially classified contingency table (Little and Rubin 1987, page 181). Under mild conditions on the non-response mechanism (for example, strictly positive and constant probabilities for each response configuration (Thibaudeau 1988)) the likelihoods in (2) and (3) are identifiable and asymptotically unimodal. Multimodality is theoretically possible for finite samples, but it does not appear to occur in the cases studied in the paper, where the proportions of unreported items are small.

5. FINDING THE MLE AND DERIVING MEASURES FOR THE NON-RESPONSE ERROR

In this section, we recall how to compute \hat{P} , the MLE of P , and we derive measures of errors for $A(f)$ and another predictor $\hat{S}(s)$, which we term the "MLE" of the expected value of $S(s)$, which is the actual count of householders in state s at the tract level. An error measure for $\hat{S}(s)$ will be useful in section 7 to evaluate the imputation results

obtained with $A(f)$ relative to those with the HSHD. We compute \hat{P} by maximizing (3), in terms of Q , with the EM algorithm. Because of the factorization described in section 4, this maximum also yields \hat{P} .

To derive measures of error in predicting $S(s)$ for a given s , consider all the triples of the form (t, Z, v) in (1) that are observed in the sample (*i.e.*, the tract) for which it is possible, but due to item non-response it is not known, that one or more householders corresponding to such a triple are in state s . Let $\Lambda(s)$ be the number of such triples. We index these triples with $\lambda = 1, \dots, \Lambda(s)$. Let $\delta(\lambda)$ be the number of householders corresponding to triple λ , and let $\rho_\lambda(s)$ be the probability that such a householder is indeed in state s , where $\rho_\lambda(s)$ is derived from P . Let $\Delta(s, \lambda)$ be the unknown number of householders who are indeed in state s among the $\delta(\lambda)$ candidates. Based on our model we have $S(s) = S_{obs}(s) + \sum_{\lambda=1}^{\Lambda(s)} \Delta(s, \lambda)$, where $S_{obs}(s)$ is the number of householders who reported being in state s and $\Delta(s, \lambda)$ is $\text{Binomial}(\delta(\lambda), \rho_\lambda(s))$. Furthermore, let $\hat{S}(s) = S_{obs}(s) + \sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda) \hat{\rho}_\lambda(s)$, where $\hat{\rho}_\lambda(s)$ is the MLE of $\rho_\lambda(s)$. If we treat the λ 's as independent predictors, like in a regression situation, and since \hat{P} is asymptotically normal with mean P , we have the following large sample approximation for the MSE of $\hat{S}(s)$ in predicting $S(s)$.

$$E \left[\left(\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda) \hat{\rho}_\lambda(s) - \Delta(s, \lambda) \right)^2 \middle| P \right] \\ \approx V \left(\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda) \hat{\rho}_\lambda(s) \middle| P \right) + V \left(\sum_{\lambda=1}^{\Lambda(s)} \Delta(s, \lambda) \middle| P \right). \quad (4)$$

Let V_p and V_e be the first and second variances on the RHS of (4). Gelman and Rubin (1991), Larsen (1996), and Schafer (1997, page 324) introduce data augmentation Bayesian iterative proportional fitting (DABIPF) to simulate posterior and predictive distributions associated with log-linear models with data missing at random. We can use DABIPF to approximate model-consistent estimators for V_p and $V_e + V_p$ through simulations of the posterior distribution of $\sum_{\lambda=1}^{\Lambda(s)} \delta(\lambda) \rho_\lambda(s)$ and the predictive distribution of $S(s)$ respectively. Furthermore, we approximate the MSE of the demographic counts obtained imputing with $A(f)$ by adding another V_e to $V_e + V_p$ in (4) to account for the additional noise of the "dice roll" involved in $A(f)$.

6. MODELING AND SENSITIVITY ANALYSIS

6.1 A Conditional Independence Model for Sacramento

Using the notation of section 3, the householder variables in Ψ are race, origin, tenure, and sex. The categories for race are White, Black, Asian, and Other. For origin they are Hispanic and non-Hispanic. For tenure they are owner and renter. For sex they are male and female.

The neighbor variables in Ξ are race, origin, and tenure. The categories for race of the neighbor are Black and non-Black. The categories for origin and tenure of the neighbor are the same as for the householder. We design Θ in (3), by selecting interactions between the variables in Ψ and Ξ . To ensure equivalence between (2) and (3), we select the interactions of all orders between the variables in Ξ . We attempt to maintain through the imputations the correlation between successive householders in f in terms of each item in Ξ . Thus we include each interaction associating an item in Ξ to the corresponding item in Ψ . We complete the model by selecting consistency associations: We include the six interactions representing the associations involving a pair of items in Ψ . The resulting contingency table has 256 cells, and the log-linear model has thirty free parameters.

This model leads to a conditional independence transition structure. For example, conditional on the race of the closest complete neighbor, the race of the householder is independent of the tenure of the closest complete neighbor. Conditional independence allows us to combine neighbor information obtained from multiple neighbors to produce a synthetic closest complete neighbor. This approach ensures that we can use all the information available from the closest neighbor, even if he is not complete. With this approach, the correlation structure among the items of the householder is maintained whenever only one item per householder is imputed. In Sacramento, among 138,271 householders, approximately 0.1% did not report sex, 3.5% did not report race, 2.9% did not report origin, and 7.6% did not report tenure. Furthermore, race and origin are missing jointly for 0.49% of the householders, race and tenure 0.48%, origin and tenure 0.69%. Given these low rates of jointly missing items, we expect our model to do well.

6.2 Sensitivity Analysis and Evaluation

In section 7 we use the standard error of the predictive distribution of $S(s)$ to approximate $\sqrt{V_e + V_p}$, the error of $\hat{S}(s)$ in predicting $S(s)$, as derived in (4), and we assume asymptotic normality of $\hat{S}(s) - S(s)$. The accuracy of this approximation depends on the accuracy of the approximation of the distribution of the MLE \hat{P} with the posterior distribution of P . This later approximation is accurate asymptotically when the model holds, but we still need to verify the extent to which this asymptotic result is applicable when the sample is finite. To do so we examine the sensitivity of the posterior distribution of P under prior changes. A low sensitivity implies that the posterior distribution of P is a good approximation of the distribution of \hat{P} . We focus on the posterior distribution for the conditional probability that origin is Hispanic, conditional on each race. An increase of .1 in the value of α , the prior parameter of the constrained Dirichlet family (Schafer 1997, page 346), which is the natural family for (3), is equivalent to observing three additional Hispanics and three additional Non-Hispanics of each race. Table 3 gives the posterior

modes and standard deviations (SD) of the posterior density of the conditional probability that origin is Hispanic given each race, for four choices of α , for a specific tract X. Figure 1 shows the posterior of the conditional probability given race is White. This posterior is stable under prior disturbances and we expect it to give a good approximation for the distribution of the corresponding MLE. On the other hand, Figure 2, which shows the posterior of the conditional probability given race is Black, displays a high sensitivity, suggesting that our proposed asymptotic approximation is less accurate in this case. This is not surprising in light of the facts that, for Blacks, the MLE of the conditional probability is close to 0 and the domain (race) size is smaller (among the 1,583 householders in tract X, there are 1,087 Whites, 179 Blacks, 56 Asians, 172 Others, while 89 did not report race). In the next section we focus on cases where the conditional probabilities are not near 0 or 1, and the size of the domain is large. We retain the choice $\alpha = .01$ for the prior, which is approximately Jeffrey's prior on the marginal conditional probabilities that define the model. It is beyond the scope of the paper to address the difficulties when the domain is small and/or the MLE is near 0/1.

Table 3

MLE, Posterior Mode (approximate), and Standard Deviation for the Conditional Probabilities of Origin Being Hispanic Given Race for Four Choices of Prior Distribution

Race	MLE	Mode $\alpha=.01$	S.D. Mode $\alpha=.01$	Mode $\alpha=.1$	S.D. Mode $\alpha=.1$	Mode $\alpha=.5$	S.D. Mode $\alpha=.5$	Mode $\alpha=1$	S.D. Mode $\alpha=1$
White	.1784	.178	.01195	.184	.01247	.180	.01219	.188	.01186
Black	.07428	.0690	.02272	.081	.02330	.120	.02428	.160	.02782
Asian	.09113	.105	.04086	.108	.04550	.195	.04881	.276	.04952
Other	.9662	.966	.01171	.964	.01347	.950	.01495	.930	.01666

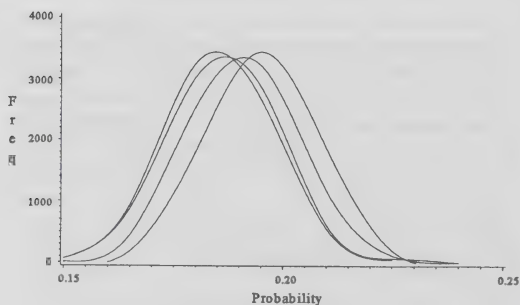


Figure 1. Posterior Distribution
Prob. Origin is Hispanic – White Householder

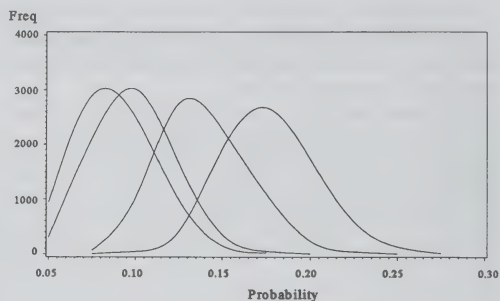


Figure 2. Posterior Distribution
Prob. Origin is Hispanic – Black Householder

7. RESULTS FOR THE SACRAMENTO DRESS REHEARSAL

Table 4 gives count estimates at the level of Sacramento derived with $A(f)$ based on the model of section 6.1 fitted for each of the 102 tracts, as well as count estimates obtained with the HSHD. Table 4 also gives error measurements based on a sequence of 2000 DABIPF iterations with 2000 burn-in iterations, for each of the 102 tracts in Sacramento (see appendix A for convergence), serving to approximate $\sqrt{V_e + V_p}$ derived from (4). We call $\sqrt{V_e + V_p}$ the prediction error of the MLE. We estimate $\sqrt{V_e}$ separately by “rolling dice” loaded with the MLE. We call $\sqrt{V_e}$ the model residual error. We use $\sqrt{2V_e + V_p}$, which we call the total imputation error, to express the error of $A(f)$ in estimating the true count. If we assume $\hat{S}(s)$ is positively correlated with the HSHD, the prediction error of the MLE can be used as an upper bound for the standard error of the distance between the count estimates corresponding to the MLE and the HSHD. For the Black owners, this distance is severely incompatible with the hypothesis that the MLE and the HSHD have the same expectation. This is no surprise in light of the results of section 2.

Interestingly, the results of table 4 can serve to improve the performance of the HSHD. Since tenure is unreported twice as often as race, our results for the Black owners suggest improving the HSHD by including race as a class variable for the imputation of tenure with the HSHD. Table 5 shows results obtained with this re-engineered HSHD, and exchangeability of tenure between nearest neighbors based on this new post-stratification is more plausible than for the original scheme.

Table 4
Population Counts and Uncertainty Measures for Sacramento

	Imputed Count With HSHD	Imputed Count With Model	MLE of the Expected Count	Model Residual Error	Prediction Error of the MLE	Total Imputation Error
All	138,271	138,271	138,271.0	0.0	0.0	0.0
White	89,032	88,914	88,927.7	31.5	35.2	47.2
Black	19,962	19,943	19,952.9	14.9	16.5	22.3
Asian	17,405	17,421	17,426.2	14.0	14.9	20.5
Other	11,872	11,993	11,964.1	29.8	33.5	44.8
Hispanic	21,024	21,050	21,038.1	10.3	10.6	14.7
Non-Hispanic	117,247	117,221	117,232.8	10.3	10.6	14.7
Owner	70,054	70,022	70,026.3	42.8	43.3	60.9
Renter	68,217	68,249	68,244.7	42.8	43.3	60.9
White Hispanic	9,068	8,972	8,991.1	29.9	33.6	45.0
White Non-Hispanic	79,964	79,942	79,936.6	15.4	15.7	22.0
Black Hispanic	605	612	608.6	11.0	12.6	16.7
Black Non-Hispanic	19,357	19,331	19,344.3	10.8	10.7	15.2
Asian Hispanic	518	515	516.5	10.0	11.5	15.2
Asian Non-Hispanic	16,887	16,906	16,909.7	10.4	10.3	14.6
Other Hispanic	10,833	10,951	10,921.9	29.7	33.3	44.6
Other Non-Hispanic	1,039	1,042	1,042.3	3.5	3.4	4.9
White Owner	47,722	47,767	47,770.5	37.8	41.3	56.0
White Renter	41,310	41,147	41,157.3	39.0	41.4	56.9
Black Owner	7,661	7,538	7,542.3	19.6	20.7	28.5
Black Renter	12,301	12,405	12,410.6	21.1	22.5	30.8
Asian Owner	9,810	9,853	9,872.8	18.4	18.6	26.1
Asian Renter	7,595	7,568	7,553.4	18.2	18.8	26.1
Other Owner	4,861	4,864	4,840.7	24.4	28.2	37.3
Other Renter	7,011	7,129	7,123.4	25.4	28.6	38.2
Hispanic Owner	9,409	9,434	9,402.2	19.5	20.9	28.6
Hispanic Renter	11,615	11,616	11,629.9	20.1	21.4	29.4
Non-Hispanic Owner	60,645	60,588	60,618.0	38.9	39.4	55.4
Non- Hispanic Renter	56,602	56,633	56,614.8	38.7	39.6	55.4

Table 5
HSHD with Race as an Additional Class Variable

	Imputed Count with HSHD	Imputed count with HSHD re- engineered with Race as a Class Variable	Imputed Count with Model	MLE of the Expected Count	Prediction Error of the MLE
White owner	47,722	47,687	47,767	47,770.5	41.3
Black Owner	7,661	7,573	7,538	7,542.3	20.7
Asian Owner	9,810	9,851	9,853	9,872.8	18.6
Other Owner	4,861	4,840	4,864	4,840.7	28.2
Owner	70,054	69,951	70,022	70,026.3	43.3

8. CONCLUSION

In section 2 we have shown that the HSHD may fail to retrieve exchangeable householders, producing a bias relative to a situation where exchangeability holds. As more evidence that $A(f)$ partly corrects this relative bias, we compare the observed and the imputed cross-product ratios (Bishop, Fienberg and Holland 1975, page 14) between two races (Black, White) and the two tenures. We look at the cross product ratio involving:

1. Only observed householders.
2. Householders with tenure imputed with the HSHD.
3. Householders with tenure imputed with $A(f)$.

There are 73 tracts where all these cross-product ratios can be measured. 2. The HSHD produces cross-product ratios smaller than those observed for 53 tracts. $A(f)$ displays more symmetry as it produces cross-product ratios smaller than observed only for 43 tracts. A sign test confirms that $A(f)(p = .064)$ is more in sync with the observations than the HSHD ($p = .0001$).

In general, we expect the HSHD to give good count estimates when the householders tend to geographically coalesce by domain of study. But difficulties arise in a situation where domains of study exhibiting substantial distributional dissimilarities are geographically integrated. In such a situation, implementing the HSHD requires accurate parsing of the class variables. Frugality is tantamount when specifying class variables, but at the same time the price to pay for omitting a crucial variable can be substantial. Thus the designer of the HSHD has little room for error. By contrast, although model misspecification certainly remains a danger, the user of $A(f)$ has more freedom to posit several domain partitions without impeding on the ability of $A(f)$ to adjust the imputations for the local geographical character, based on information from the closest complete neighbor. $A(f)$ will be useful to impute categorical measurements when the impact of the relevant domain partitions on the measurements is not known a priori, and some of the relevant domains may define small subpopulations dispersed within the entire population. Then, based on policy considerations, $A(f)$ can be applied directly, or to help parse the class variables of the HSHD, as we did in section 7.

A referee notes that a comparison with a procedure based on an unbiased sample, building on the method of Zanutto and Zaslavsky (1995a,b), would be a defining test for $A(f)$. This procedure would require collecting information from the item non-respondents on a scale sufficiently large to ensure bias detection, and we should take advantage of any such opportunity to perform a test of this type. Unfortunately, because of limited resources, samples containing this information are seldom collected. Nevertheless, we are hopeful that the analysis of the returns from Census 2000 aided with procedural information can provide new insights on the reliability of $A(f)$.

ACKNOWLEDGEMENTS

The author is indebted to William Winkler for his guidance. The author is grateful to two referees for their discernment, to Eric Slud, Don Malec and Joseph Schafer for essential discussions, and to Andrew Gelman and Don Rubin for providing their unpublished paper. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

APPENDIX A – CONVERGENCE OF DABIPF

We ran two chains of 8,000 iterations each, with over-dispersed starting points, for the case $\alpha = 0.01$, for tract X. We computed $\sqrt{\hat{R}}$ (Gelman and Rubin 1992) for $Q(s, t)$ in (3), for sequences of 1,000, 2,000, and 4,000 iterations, after burn-in lags of 1,000, 2,000, and 4,000 iterations respectively. After 2,000 iterations, with 2,000 burn-in iterations, we observed that $\sqrt{\hat{R}} \leq 1.010$ in all studied cases, including those in table 3. We think this level of accuracy is acceptable for approximating modes and variances.

REFERENCES

- AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley-Interscience.
- BESAG, J. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society, B*. 36, 2.
- BISHOP, Y. M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press.
- CHEN, J., and SHAO, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 365-369.
- CHEN, J., and SHAO, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*. 16, 2.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd Edition. Wiley.
- DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 39, 1-22.
- FAY, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 227-232.
- FAY, R.E. (1999). Theory and application of nearest neighbor imputation in census 2000. *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 112-121.
- FAY, R.E., and TOWN, M.K. (1998). Variance estimation for the 1998 census dress rehearsal. *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 605-610.
- GELMAN, A., and RUBIN, D.B. (1991). Simulating the Posterior Distribution of Loglinear Contingency Table Models. Unpublished Technical Report, Harvard University.
- GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*. 7, 4.
- KOSTANICH, D.L. (1999). DSSD Census 2000 Dress Rehearsal Memorandum Series #A, US Bureau of the Census.
- KOVAR, J.G., and WHITRIDGE, P.J. (1995). Imputation of Business Survey Data. *Business Survey Methods*, (Eds. Cox, D. Binder, Chinnappa, Christianson, M. Colledge and Kott). Wiley.
- LARSEN, M.D. (1996). *Bayesian Approaches to Finite Mixture Models*. Doctoral Dissertation, Department of Statistics, Harvard University.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley.

- MENG, X.M. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 9, 4.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 79, 4.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 91, 434.
- SANDE, I.G. (1981). Imputation in surveys: coping with reality. *Survey Methodology*. 7, 21-43.
- SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- THIBAudeau, Y. (1988). *Approximating the Moments of a Multimodal Posterior Distribution with the Method of Laplace*. Doctoral Dissertation, Department of Statistics, Carnegie Mellon University.
- TREAT, J.B. (1994). *Summary of the 1990 Census Imputation Procedures for the 100 % Population and Housing Items*. DSSD REX Memorandum Series BB-11, US Bureau of the Census.
- WILSON, E.B. (1998). Communication to Dan E. Philip. Housing and Household Economics Statistics Division, US Bureau of the Census.
- ZANUTTO, E., and ZASLAVSKY, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. *Proceedings for the Section on Survey Research Methods*, American Statistical Association. 608-613.
- ZANUTTO, E., and ZASLAVSKY, A.M. (1995b). Models for imputing nonsample households with sampled nonresponse follow-up. *Proceedings of the Annual Research Conference*, U.S. Department of Commerce, Bureau of the Census. 673-686.

A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas

BALGOBIN NANDRAM, GEUNSHIK HAN and JAI WON CHOI¹

ABSTRACT

The analysis of survey data from different geographical areas, where the data from each area are polychotomous, can be easily performed using hierarchical Bayesian models even if there are small cell counts in some of these areas. However, there are difficulties when the survey data have missing information in the form of nonresponse especially when the characteristics of the respondents differ from the nonrespondents. We use the selection approach for estimation when there are nonrespondents because it permits inference for all the parameters. Specifically, we describe a hierarchical Bayesian model to analyze multinomial nonignorable nonresponse data from different geographical areas, some of them can be small. For the model, we use a Dirichlet prior density for the multinomial probabilities and a beta prior density for the response probabilities. This permits a "borrowing of strength" of the data from larger areas to improve the reliability in the estimates of the model parameters corresponding to the smaller areas. Because the joint posterior density of all the parameters is complex, inference is sampling based and Markov chain Monte Carlo methods are used. We apply our method to provide an analysis of body mass index (BMI) data from the third National Health and Nutrition Examination Survey (NHANES III). For simplicity, the BMI is categorized into three natural levels, and this is done for each of eight age-race-sex domains and thirty-four counties. We assess the performance of our model using the NHANES III data and simulated examples, which show our model works reasonably well.

KEY WORDS: Latent variable; Metropolis-Hastings sampler; Nonignorable nonresponse; Selection approach; Small area.

1. INTRODUCTION

The nonresponse rates in many surveys have been increasing steadily (De Heer 1999; Groves and Couper 1998), making the nonresponse problem more important. For many surveys the responses are polychotomous. For example, in the third National Health and Nutrition Examination Survey (NHANES III), we can estimate the proportions of persons belonging to three levels of body mass index (BMI), although BMI is a continuous variable. The purpose of this paper is to describe a new hierarchical Bayesian model to study nonignorable multinomial nonresponse for small areas, and to apply it to the NHANES III BMI data.

Rubin (1987) and Little and Rubin (1987) describe two types of models which differ according to the ignorability of response. In the ignorable nonresponse model the distribution of the variable of interest for a respondent is the same as the distribution of that variable for a nonrespondent with the same values of the covariates. In addition, the parameters in the distributions of the variable and response must be distinct (see Rubin 1976). All other nonresponse models are nonignorable. We use both ignorable and nonignorable nonresponse models for our data because there are no nonrespondents for some domains.

Crawford, Johnson and Laird (1993) used nonignorable nonresponse models to analyze data from the Harvard Medical Practice Survey. Stasny, Kadane, and Fritsch

(1998) used a Bayesian hierarchical model for the probabilities of voting guilty or not on a particular trial when the views of nonrespondents differ from those of respondents in various death-penalty beliefs. Park and Brown (1994) used a pseudo-Bayesian method (Baker and Laird 1988), and Park (1998) applied a method in which prior observations are assigned to both observed and unobserved cells to estimate the missing cells of a multi-way categorical table under nonignorable nonresponse. Our approach differs from these authors. We describe small area estimation for multinomial data, and we use Markov chain Monte Carlo methods to implement the methodology. This permits the inclusion of all sources of variability in our models.

There are two approaches to model nonresponse. The selection approach is used for the hypothetical complete data, and a nonresponse model is added conditional on the hypothetical data. This approach was developed to study sample selection problems (e.g., Heckman 1976 and Olson 1980). In the pattern mixture approach the respondents and the nonrespondents are modeled separately, and the final answer is obtained by a probabilistic mixture of the two. We use the selection approach for our problem.

Stasny (1991) used an empirical Bayes model to study victimization in the National Crime Survey, and she followed the selection approach. This analysis pools binomial data from several domains, and some of them have small counts. Essentially this is an exercise in small area estimation. A related method was presented by Albert and

¹ Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA (balnan@wpi.edu); Geunshik Han, Division of Computer and Information Science, Hanshin University, Osan, Korea. (gshan@hucc.hanshin.ac.kr); Jai Won Choi, National Center for Health Statistics, Room 915, 6525 Belcrest Road, Hyattsville, MD 20782, USA, (jwc7@cdc.gov).

Gupta (1985), who used an approximation to obtain a Bayesian approach for a population with a single domain (see also Kaufman and King 1973). That is, unlike Stasny (1991), these latter authors did not perform small area estimation, and their analysis in a single domain do not use data from other domains.

Since the Bayesian approach can incorporate other information about nonrespondents, the Bayesian method is appropriate for the analysis of nonignorable nonresponse (Little and Rubin 1987 and Rubin 1987). However the main difficulty is how to describe the relationship between the respondents and nonrespondents. Using the selection approach within the framework of Bayes empirical Bayes (see Deely and Lindley 1981), Stasny (1991) estimated the hyper-parameters by maximum likelihood methods and then assumed them known, thereby suppressing some variability. We extend this approach in two directions.

First, we consider multinomial data obtained independently from several geographical areas. It is worthy to note that Basu and Pereira (1982) considered multinomial nonresponse data from a single domain using a multinomial Dirichlet model when the hyper-parameters are assumed known. Recently, Forster and Smith (1998) used graphical multinomial Dirichlet log-linear models to analyze data from the panel survey in British general election. Again the hyper-parameters are assumed known, and a model with a single domain is used. Secondly, we obtain a full Bayesian approach for multinomial nonignorable nonresponse data from several areas. We do not estimate the hyper-parameters using the data.

As a summary, we develop a multinomial nonignorable nonresponse model which is used for pooling data over many small areas, and we note that it can be used in other applications. The rest of the paper is organized as follows. In section 2 we describe the NHANES III. In section 3 we discuss the Bayesian model for nonignorable nonresponse. In particular, a three-stage Bayesian hierarchical multinomial model is applied to the NHANES III data to investigate the nonresponse problem. In section 4 we describe an analysis of the NHANES III data in which we include a regression analysis to combine all the age-race-sex domains. In section 5 we describe a simulation study to assess the performance of our model. Finally, section 6 has the conclusion.

2. NHANES III DATA AND NONRESPONSE

The NHANES III is one of the periodic surveys used to assess an aspect of health of the U.S. population (National Center for Health Statistics 1994). Our research is motivated by nonresponse of body mass index (BMI) in the NHANES III. The data for our illustration come from this survey, and were collected from October 1988 to September 1994. In section 2.1 we describe the actual data, and in section 2.2 we describe the data we analyze.

2.1 NHANES III Data

The NHANES III consists of two parts. The first part is the interview of the sampled individuals for their personal information and the second part is the examination of those sampled. One or more persons from the sampled households were placed into a number of subgroups depending on their age, race and sex. Some subgroups were sampled at different rates. Sampled persons were asked to come to a mobile examination center (MEC) for a physical examination. Those who did not come were visited by the examiner for the same purpose. Details of the NHANES III sample design are available (National Center for Health Statistics 1992). We incorporate design features associated with clustering in our model.

The main reasons for NHANES III nonresponse are "not interested", "no time/work conflict", "concerns/suspicious", "don't bother me" and "health reasons". The nonresponse rate of younger individuals is very high because the parents, especially older mothers of an only child, were extremely protective of their babies, and would not allow them to leave their homes for the MECs. Field workers often observe that obese persons tend to avoid the medical examination. So that nonresponse might be nonrandom and hence require some special attention.

NHANES III data are adjusted by multistage ratio weightings for the data to be consistent with the population (Mohadjer, Bell and Waksberg 1994). The ratio is the proportion of persons in the sample to the number of persons who completed interview and examination. Weighting with nonresponse ratio is one of these stages. In nonresponse ratio estimation, the proportions of nonrespondents in the multinomial cells are the same as those for the respondents (*i.e.*, ignorable nonresponse). In this case since the proportions are of interest, no adjustment is required. Clearly, this ratio estimation can be incorrect when these two groups are different. Therefore there is a need to consider the adjustment by a method other than ratio adjustment. In this paper we investigate a Bayesian method as an alternative to ratio weighting for nonignorable nonresponse.

NHANES III nonresponse also occurs at several levels in the survey: interview and examination. The interview nonresponse arises from sample individuals who did not respond for the interview. Some of those who were already interviewed did not come to the MEC, missing all or part of the examinations. In this paper, our population consists of those individuals who would have agreed to take the physical examination in the MECs. Thus, nonrespondents are those individuals who agreed to take the physical examination, and did not show up at the MECs. More specifically, since we are considering item response, the nonrespondents are those individuals who agreed to come to the MECs and their heights and/or weights were not measured.

Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin (1996) attempted a comprehensive multiple imputation project on the NHANES III data for many variables. The

purpose was to impute the nonresponse data to provide several data sets for public use. Unfortunately, one of the limitations of the project was that “the procedure used to create missingness corresponds to a purely ignorable mechanism; the simulation provides no information on the impact of possible deviations from ignorable nonresponse.” Another limitation is that the procedure did not include geographical clustering. Our purpose is different; we do not provide imputed public-use data.

2.2 Data Used for Illustration

Our data have two age groups (younger than 45 years, 45-, and 45 years or older, 45+), two race groups (white and non-white) and, of course, two groups for sex (male and female). Thus, there are eight age-race-sex domains.

One of the variables of interest in the NHANES III is BMI, an index of weight adjusted for height (Kg / m^2), that broadly categorizes obesity within age-race-sex groups (Kuczmarski, Carrol, Flegal and Troiano 1997) as low body fat (level 1: BMI < 20), healthy body fat (level 2: $20 \leq BMI < 25$), hefty or unhealthy (level 3: BMI ≥ 25). We use this broad classification for each of the eight age-race-sex groups.

Rather than a categorical data analysis, one can also provide an analysis that treats BMI as a continuous variable. While some information is lost by discretizing the BMI values, an analysis using continuous models for BMI will also be approximate and there is a need to search for an appropriate transformation. In the final analysis, a doctor only needs to know what proportions of the public belong to different levels of BMI, so he or she can tell his patient’s standing in obesity.

The analysis of BMI data using categorical data methods is not uncommon. For example, Malec, Davis and Cao (1999) described a Bayes empirical Bayes analysis of the NHANES III data. They classified an individual older than 20 years as normal if her/his BMI is below a certain gender specific threshold. This is an application of a Bayesian analysis of binary data. However, their classification is somewhat restricted (see Kuczmarski *et al.* 1997). By considering multinomial data, we have generalized the analysis of Malec *et al.* (1999). In fact, they did not provide a nonignorable nonresponse model.

Unlike Schafer *et al.* (1996), we include clustering at the county level, although there is a need to include clustering at the household level. For the complete data there are 6,440 households. Of these households 52.1% contributed one person to the sample, 22.5% two persons, and 21.4% at least three persons. We have calculated the correlation coefficient for the BMI values based on pairing the members within households (see Rao 1973 page 199). It is 0.19 which indicates that as a first approximation the clustering within households can be ignored.

Table 1 shows the number of respondents for each BMI level for each age-race-sex domain and 34 counties (population at least 500,000). The pattern of respondents

differs greatly by age. The nonresponse rate for the older group (45+) is negligible. Therefore the main concern about nonresponse must be given to the younger group (45-). There is also higher response rate among females than males. We note that the selection procedure is not random over the single population of males and females.

Table 1
Number of individuals in each BMI level and number of nonrespondents (Non) by age, race and sex over all 34 counties

Age	Race	Sex	BMI			
			1	2	3	Non
45-	W	M	1,098	651	597	558
		F	845	434	380	233
	B	M	1,198	713	665	574
		F	745	463	524	214
45+	W	M	46	439	1,014	3
		F	51	223	365	4
	B	M	79	470	942	8
		F	48	169	552	6

Note: BMI (1=less than 20; 2 = at least 20 and smaller than 25; 3 = greater than 25)
Age (Younger than 45 years = 45-; 45 years or older = 45+)
Race (White = W; all others = B)
Sex (Male = M; Female = F)

Table 2
Number of individuals in each BMI level and number of nonrespondents (Non) for eight examples (Ex) of small age-race-sex domains from different counties

Ex	Age	Race	Sex	BMI Level			Non
				1	2	3	
1	45-	W	M	1	3	1	14
2			F	3	4	1	0
3		B	M	5	5	6	10
4			F	3	1	1	1
5	45+	W	M	1	2	6	0
6			F	1	3	4	0
7		B	M	3	3	5	0
8			F	2	0	1	1

Note: BMI (1=less than 20; 2 = at least 20 and smaller than 25; 3 = greater than 25)
Age (Younger than 45 years = 45-; 45 years or older = 45+)
Race (White = W; all others = B)
Sex (Male = M; Female = F)

One important aspect of our work is on small area estimation. Because we consider inference for each age-race-sex domain separately over the the geographical areas (counties), the samples from some of these areas can be very small. Thus, small area estimation techniques are required to estimate the parameters corresponding to these smaller areas. Specifically, we need to “borrow strength” from the larger areas to make the estimates for the smaller areas more reliable. Table 2 presents eight examples to show the need for small area techniques. We have selected eight counties that have small domains; all the cell counts are at most 6 and many of them are as small as 1 (one of

them is 0 for 45+). We will present overall estimates and the estimates for the first four examples (45-). Note that in comparison to the cell counts, the nonrespondents are large for two of them (14 and 10 nonrespondents).

We note that the purpose is not a comprehensive analysis of the NHANES III data although it forms an approximate analysis for these data. Our method is general enough to analyze multinomial nonresponse data from many areas, some of which can be small. It is for these small areas that we develop this modeling technique. Thus, in this paper we use the NHANES III data to illustrate our method.

Our method considers each domain separately with a “borrowing of strength” across the 34 areas (counties) to analyze the BMI data. Thus, there are eight separate analyses, each with 34 areas, and some of them are small. We use a hierarchical multinomial nonresponse model to analyze data of this form. The small cell counts, substantial nonrespondents and multinomial data make the methodology much more practical. Our methodology is also extended to incorporate all the domains simultaneously through logistic models.

3. METHODOLOGY FOR HIERARCHICAL MULTINOMIAL MODEL

We propose a model for each of the eight age-race-sex domains but for all counties taken simultaneously. However, the models fall into two broad classes. We will use a nonignorable nonresponse model for the younger group and an ignorable nonresponse model for the older group since the nonresponse rate for the older group is negligible. Of course, it is worthwhile to compare the ignorable nonresponse model and the nonignorable nonresponse model for the younger group. We will show how to combine the groups later using logistic regression, although this is not the key issue of this paper.

For each age-race-sex group, the k^{th} individual in the i^{th} county belongs to one of J BMI levels. Then for the k^{th} individual in i^{th} county, the characteristic variable at the j^{th} BMI level is defined as follows,

$$\mathbf{x}_{ik} = (x_{i1k}, \dots, x_{ijk}, \dots, x_{iJk})', \quad i = 1, \dots, c, \quad k = 1, \dots, n_i,$$

where each $x_{ijk} = 0$ or 1 , $j = 1, \dots, J$, and $\sum_{j=1}^J x_{ijk} = 1$. The response variable, y_{ijk} , is defined for each age-race-sex domain

$$y_{ijk} = \begin{cases} 1, & \text{if individual } k \text{ belonging to BMI level } j \text{ in county } i \text{ responded} \\ 0, & \text{if individual } k \text{ belonging to BMI level } j \text{ in county } i \text{ did not respond.} \end{cases}$$

We use a probabilistic structure to model the \mathbf{x}_{ik} and y_{ijk} . In our application, there are $c = 34$ counties and $J = 3$ BMI levels.

3.1 Ignorable and Nonignorable Nonresponse Models

For both ignorable and the nonignorable nonresponse models, we have

$$\mathbf{x}_{ik} | \mathbf{p}_i \stackrel{\text{iid}}{\sim} \text{Multinomial}(1, \mathbf{p}_i) \quad (1)$$

where p_{ij} is the probability that an individual in the i^{th} county belongs the j^{th} BMI level. Next, we describe the remaining portions of the ignorable and the nonignorable models.

First, we describe the ignorable nonresponse model. Let π_i denote the probability that an individual within the i^{th} county responds (*i.e.*, the probability of responding depends only on the county). Then, we assume that

$$y_{ijk} | \pi_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\pi_i). \quad (2)$$

At the second stage, letting $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \dots, \mu_{1J})'$, we take

$$\mathbf{p}_i | \boldsymbol{\mu}_1, \boldsymbol{\tau}_1 \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1), \quad (3)$$

$$\pi_i | \mu_{21}, \tau_{21} \stackrel{\text{iid}}{\sim} \text{Beta}(\mu_{21}, \tau_{21}, (1 - \mu_{21})\tau_{21}) \quad (4)$$

where

$$p(\mathbf{p}_i | \boldsymbol{\mu}_1, \boldsymbol{\tau}_1) = \prod_{j=1}^J p_{ij}^{\mu_{1j}\tau_{1j}-1} / D(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1), \quad 0 < p_{ij} < 1, \sum_{j=1}^J p_{ij} = 1$$

and

$$D(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1) = \prod_{j=1}^J \Gamma(\mu_{1j}\tau_{1j}) / \Gamma(\tau_{1j}), \quad 0 < \mu_{1j} < 1, \sum_{j=1}^J \mu_{1j} = 1.$$

The components of $\boldsymbol{\mu}_1$ are the prior means of the corresponding components of the \mathbf{p}_i , and $\boldsymbol{\tau}_1$ can be interpreted as a prior sample size. Similar interpretations can be given for μ_{21} and τ_{21} for π_i . Thus, assumption (3) expresses similarity among the cell proportions \mathbf{p}_i and (4) expresses similarity among the response probabilities π_i . It is this structure that causes the “borrowing of strength” across the c counties.

Second, we describe the nonignorable nonresponse model. Let π_{ij} denote the probability that an individual within the i^{th} county responds in the j^{th} BMI level (*i.e.*, the probability of responding depends not only on the county but also on the BMI level). Then, we assume that

$$y_{ijk} | \{x_{ik} = (x_{i1k}, \dots, x_{iJk}), \pi_{ij}\} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\pi_{ij}) \quad (5)$$

where $x_{ijk} = 1$, $x_{i'j'k} = 0$, $j \neq j'$ for $j, j' = 1, 2, \dots, J$. Letting $\boldsymbol{\mu}_3 = (\mu_{31}, \mu_{32}, \dots, \mu_{3J})'$, at the second stage we also take

$$\mathbf{p}_i | \boldsymbol{\mu}_3, \boldsymbol{\tau}_3 \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3) \quad (6)$$

and

$$\pi_{ij} | \mu_{4j}, \tau_{4j} \stackrel{\text{iid}}{\sim} \text{Beta}(\mu_{4j}\tau_{4j}, (1 - \mu_{4j})\tau_{4j}), \quad j = 1, \dots, J. \quad (7)$$

Like the assumptions in (3) and (4), the assumptions in (6) and (7) express similarity among the counties. We note that the response parameters π_{ij} are weakly identifiable (*i.e.*, unreliable estimates). However, the selection model works to our advantage, because the joint density of \mathbf{x}_{ik} and $\mathbf{y}_{ik} = (y_{i1k}, \dots, y_{iJk})'$ connects the p_{ij} and π_{ij} . In fact, this is an advantage over the pattern mixture approach.

To ensure a full Bayesian analysis, at the third stage we take the prior densities for the hyper-parameters as follows. For the ignorable nonresponse model, the prior densities are

$$\mu_1 \sim \text{Dirichlet}(1, 1, \dots, 1), \mu_{21} \sim \text{Beta}(1, 1),$$

$$\tau_1 \sim \text{Gamma}(\eta_1^{(0)}, v_1^{(0)}) \text{ and } \tau_{21} \sim \text{Gamma}(\eta_{21}^{(0)}, v_{21}^{(0)}),$$

where (letting t denote either τ_1 or τ_{21} , a either $\eta_1^{(0)}$ or $\eta_{21}^{(0)}$, and b either $v_1^{(0)}$ or $v_{21}^{(0)}$) $\tau \sim \text{Gamma}(a, b)$ means that $f(t) = b a^t t^{a-1} e^{-bt} / \Gamma(a)$, $t > 0$ and $f(t) = 0$ otherwise. The hyper-parameters $\eta_1^{(0)}$, $v_1^{(0)}$, $\eta_{21}^{(0)}$ and $v_{21}^{(0)}$ are to be specified. The corresponding part of the nonignorable nonresponse model is

$$\mu_3 \sim \text{Dirichlet}(1, 1, \dots, 1), \mu_{4j} \sim \text{Beta}(1, 1),$$

$$\tau_3 \sim \text{Gamma}(\eta_3^{(0)}, v_3^{(0)}) \text{ and}$$

$$\tau_{4j} \sim \text{Gamma}(\eta_{4j}^{(0)}, v_{4j}^{(0)}), j = 1, \dots, J.$$

Again, the hyper-parameters $\eta_3^{(0)}$, $v_3^{(0)}$, $\eta_{4j}^{(0)}$, $v_{4j}^{(0)}$, $j = 1, \dots, J$, are to be specified. It is possible to use other prior densities such as shrinkage priors, but it is likely that these will provide similar inference as our sensitivity analysis indicates in section 4.

It is an attractive property of the hierarchical model that it introduces correlation among the variables. For example, in our application (1), (2), (3) and (4) make the $(\mathbf{x}_{ij}, y_{ij})$ equi-correlated across the individuals within the i^{th} area. This is the clustering effect within the areas. Such an effect can be obtained directly, but it will not be as simple as in a hierarchical model. A further benefit of the hierarchical model is that it takes care of extraneous variations among the areas, and this effect can be obtained directly by using random effects model. But in our case, this will loose the natural multinomial data structure.

Let r_i be the number of respondents in county i and y_{ij} the number of respondents having the j^{th} BMI level in the i^{th} county. Then r_i and y_{ij} are random variables; $n_i - r_i$ is the number of nonrespondents. Since the number of nonrespondents at the j^{th} BMI level is unknown, we denote them by the latent variables z_{ij} (see the tree diagram in Figure 1). If we can tell what the z_{ij} are, our nonresponse problem will be solved. Of course, under the assumption of ignorable nonresponse, they can be estimated easily using ratio estimation. The z_{ij} are useful because under the assumption of nonignorable nonresponse they simplify the sampling based method to obtain estimates of the parameters of interest.

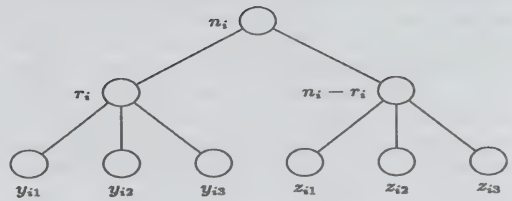


Figure 1. Latent nonignorable response tree diagram. From a sample of n_i individuals, there are r_i respondents of which y_{ij} belong to category j , $j = 1, 2, 3$. Among the $(n_i - r_i)$ nonrespondents z_{ij} individuals belong to category j , where z_{ij} are latent variables.

The likelihood function for the ignorable nonresponse model is

$$f(\mathbf{y}, \mathbf{r} | \mathbf{p}, \boldsymbol{\pi}) = \prod_{i=1}^c \left\{ \binom{n_i}{r_i} \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i} \right\} \\ \times \prod_{i=1}^c \left\{ \binom{r_i}{y_{i1}, \dots, y_{iJ}} \prod_{j=1}^J \{ p_{ij}^{y_{ij} + n_i - r_i} \} \right\}.$$

Here the likelihood function has two distinct parts, one for p_{ij} and the other for the π_i . Using Bayes' theorem the joint posterior density of all the parameters is

$$f(\mathbf{p}, \boldsymbol{\pi}, \mu_1, \tau_1, \mu_{21}, \tau_{21} | \mathbf{y}, \mathbf{r}) \\ \propto \prod_{i=1}^c \left\{ \left\{ \prod_{j=1}^J p_{ij}^{y_{ij} + n_i - r_i} \right\} \left\{ \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i} \right\} \right. \\ \times \left\{ \prod_{j=1}^J p_{ij}^{\mu_{1j} \tau_1 - 1} \right\} / D(\mu_1, \tau_1) \\ \times \left\{ \frac{\pi_i^{\mu_{21} \tau_{21} - 1} (1 - \pi_i)^{(1 - \mu_{21}) \tau_{21} - 1}}{B(\mu_{21}, \tau_{21}, (1 - \mu_{21}), \tau_{21})} \right\} \\ \left. \times \left\{ \tau_1^{\eta_1^{(0)} - 1} \exp(-v_1^{(0)} \tau_1) \right\} \left\{ \tau_{21}^{\eta_{21}^{(0)} - 1} \exp(-v_{21}^{(0)} \tau_{21}) \right\} \right\}. \quad (8)$$

Similarly, the augmented likelihood function (*i.e.*, including the \mathbf{z}_i) for the nonignorable nonresponse model is

$$f(\mathbf{y}, \mathbf{r}, \mathbf{z} | \mathbf{p}, \boldsymbol{\pi}) = \prod_{i=1}^c \left\{ \binom{n_i}{r_i} \binom{r_i}{y_{i1}, \dots, y_{iJ}} \binom{n_i - r_i}{z_{i1}, \dots, z_{iJ}} \right\} \\ \times \prod_{j=1}^J \left\{ (\pi_{ij} p_{ij})^{y_{ij}} ((1 - \pi_{ij}) p_{ij})^{z_{ij}} \right\}$$

and using Bayes' theorem the joint posterior density of all the parameters is

$$\begin{aligned}
 & f(\mathbf{p}, \boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\mu}_3, \boldsymbol{\tau}_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4 \mid \mathbf{y}, \mathbf{r}) \\
 & \propto \prod_{i=1}^c \left\{ \left(\begin{matrix} n_i - r_i \\ z_{i1}, \dots, z_{iJ} \end{matrix} \right) \prod_{j=1}^J (\pi_{ij} p_{ij})^{y_{ij}} ((1 - \pi_{ij}) p_{ij})^{z_{ij}} \right. \\
 & \quad \times \prod_{j=1}^J p_{ij}^{\mu_{3j} \tau_{3j} - 1} / D(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3) \prod_{j=1}^J \\
 & \quad \times \left. \frac{\pi_{ij}^{\mu_{4j} \tau_{4j} - 1} (1 - \pi_{ij})^{(1 - \mu_{4j}) \tau_{4j} - 1}}{B(\mu_{4j} \tau_{4j}, (1 - \mu_{4j}) \tau_{4j})} \right\} \\
 & \times \left\{ \eta_3^{(0) - 1} \exp(-\nu_3^{(0)} \tau_3) \right\} \prod_{j=1}^J \left\{ \tau_{4j}^{(0) - 1} \exp(-\nu_{4j}^{(0)} \tau_{4j}) \right\}.
 \end{aligned}$$

We consider inference about the p_{ij} , the proportion of individuals at the j^{th} BMI level in the i^{th} county, and the probability of responding,

$$\delta_i = \sum_{j=1}^J \pi_{ij} p_{ij}, i = 1, \dots, c.$$

However, the joint posterior densities in (8) and (9) are complex, and can not be used to make inference analytically. Thus, we use a Markov chain Monte Carlo algorithm to obtain estimates of the posterior distribution of the parameters. Our method is to use a Metropolis-Hastings (MH) sampler to get samples from (8) and (9) and then to use these samples to make posterior inferences about \mathbf{p}_i and δ_i .

3.2 Computations

For the ignorable nonresponse model, it is convenient to represent the posterior density function as

$$\begin{aligned}
 & f(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\tau}_1, \boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21} \mid \mathbf{y}, \mathbf{r}) \\
 & = \prod_{i=1}^c \{ f_1(\mathbf{p}_i \mid \mathbf{y}, \mathbf{r}, \boldsymbol{\mu}_1, \boldsymbol{\tau}_1) f_2(\boldsymbol{\pi}_i \mid \mathbf{y}, \mathbf{r}, \boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21}) \} \\
 & \times f_3(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1, \boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21} \mid \mathbf{y}, \mathbf{r})
 \end{aligned}$$

where $f_1(\cdot)$ is Dirichlet density,

$$\mathbf{p}_i \mid \mathbf{y}_i, \mathbf{r}_i, \boldsymbol{\mu}_1, \boldsymbol{\tau}_1 \stackrel{\text{ind}}{\sim} D(\mathbf{y}_i + \mathbf{n}_i - \mathbf{r}_i + \boldsymbol{\mu}_1, \boldsymbol{\tau}_1),$$

$f_2(\cdot)$ is beta density,

$$\pi_i \mid \mathbf{y}_i, \mathbf{r}_i, \boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21} \stackrel{\text{ind}}{\sim} \text{Beta}(r_i + \boldsymbol{\mu}_{21} \boldsymbol{\tau}_{21}, n_i - r_i + (1 - \boldsymbol{\mu}_{21}) \boldsymbol{\tau}_{21})$$

and

$$f_3(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1, \boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21} \mid \mathbf{y}, \mathbf{r})$$

$$\begin{aligned}
 & \propto \prod_{i=1}^c \left\{ D(\mathbf{y}_i + \mathbf{n}_i - \mathbf{r}_i + \boldsymbol{\mu}_1, \boldsymbol{\tau}_1) / D(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1) \right\} p(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1) \\
 & \times \prod_{i=1}^c \left\{ \frac{B(r_i + \boldsymbol{\mu}_{21} \boldsymbol{\tau}_{21}, n_i - r_i + (1 - \boldsymbol{\mu}_{21}) \boldsymbol{\tau}_{21})}{B(\boldsymbol{\mu}_{21} \boldsymbol{\tau}_{21}, (1 - \boldsymbol{\mu}_{21}) \boldsymbol{\tau}_{21})} \right\} p(\boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21})
 \end{aligned}$$

with $p(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1)$ and $p(\boldsymbol{\mu}_{21}, \boldsymbol{\tau}_{21})$ the prior distributions. Hence, f_1 and f_2 are obtained through the Gibbs kernel, while for f_3 we use the MH algorithm (Nandram 1998).

For the nonignorable nonresponse model, it is convenient to represent the posterior density function as

$$f(\mathbf{p}, \boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\mu}_3, \boldsymbol{\tau}_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4 \mid \mathbf{y}, \mathbf{r})$$

$$\begin{aligned}
 & = \prod_{i=1}^c \left\{ \left\{ \prod_{j=1}^J f_j(\pi_{ij} \mid \mathbf{y}, \mathbf{r}, \mathbf{z}, \boldsymbol{\mu}_{4j}, \boldsymbol{\tau}_{4j}) \right\} f_{J+1}(\mathbf{p}_i \mid \mathbf{y}, \mathbf{r}, \mathbf{z}, \boldsymbol{\mu}_3, \boldsymbol{\tau}_3) \right\} \\
 & \times f_{J+2}(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4, \mathbf{z} \mid \mathbf{y}, \mathbf{r}),
 \end{aligned}$$

where $f_1(\cdot), \dots, f_J(\cdot)$ are beta densities,

$$\pi_{ij} \mid y_{ij}, r_{ij}, z_{ij}, \boldsymbol{\mu}_{4j}, \boldsymbol{\tau}_{4j} \stackrel{\text{ind}}{\sim} \text{Beta}(y_{ij} + \boldsymbol{\mu}_{4j} \boldsymbol{\tau}_{4j}, z_{ij} + (1 - \boldsymbol{\mu}_{4j}) \boldsymbol{\tau}_{4j}),$$

$f_{J+1}(\cdot)$ is a Dirichlet density,

$$\mathbf{p}_i \mid \mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\mu}_3, \boldsymbol{\tau}_3 \stackrel{\text{ind}}{\sim} D(\mathbf{y}_i + \mathbf{z}_i + \boldsymbol{\mu}_3, \boldsymbol{\tau}_3)$$

and $f_{J+2}(\cdot)$ is given by

$$f_{J+2}(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3, \boldsymbol{\mu}_4, \boldsymbol{\tau}_4, \mathbf{z} \mid \mathbf{y}, \mathbf{r})$$

$$\begin{aligned}
 & \propto \prod_{i=1}^c \left\{ \left(\begin{matrix} n_i - r_i \\ z_{i1}, \dots, z_{iJ} \end{matrix} \right) \left\{ D(\mathbf{y}_i + \mathbf{z}_i + \boldsymbol{\mu}_3, \boldsymbol{\tau}_3) / D(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3) \right\} p(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3) \right. \\
 & \quad \times \left. \prod_{j=1}^J \left\{ \frac{B(y_{ij} + \boldsymbol{\mu}_{4j} \boldsymbol{\tau}_{4j}, z_{ij} + (1 - \boldsymbol{\mu}_{4j}) \boldsymbol{\tau}_{4j})}{B(\boldsymbol{\mu}_{4j} \boldsymbol{\tau}_{4j}, (1 - \boldsymbol{\mu}_{4j}) \boldsymbol{\tau}_{4j})} \right\} \right\} p(\boldsymbol{\mu}_4, \boldsymbol{\tau}_4)
 \end{aligned}$$

with $p(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3)$ and $p(\boldsymbol{\mu}_4, \boldsymbol{\tau}_4)$ the prior distributions. Thus, f_1, \dots, f_{J+1} are obtained through the Gibbs kernel, while f_{J+2} is obtained using the MH algorithm (Nandram 1998). We obtain the latent variables z_{ij} through one of the conditional posterior densities of the MH algorithm. A sketch of the procedure is given in Appendix 1.

We drew 5,500 iterates, threw out the first 500, and took every fifth (obtained by trace plots). This strategy was satisfactory to wash out the autocorrelation among the iterates and to have good jumping probabilities (0.25-0.50) for the Metropolis steps. For the computation, first we set

the hyper-parameters $\eta_1^{(0)}, v_1^{(0)}, \eta_{21}^{(0)}, v_{21}^{(0)}, \eta_3^{(0)}, v_3^{(0)}, \eta_{4j}^{(0)}, v_{4j}^{(0)}$, $j = 1, \dots, J$ equal to 0. Then we ran our MH algorithm to obtain posterior samples of $\tau_1, \tau_{21}, \tau_3$ and τ_{4j} , $j = 1, \dots, J$. To ensure proper posterior densities, we estimate $\eta_1^{(0)}, v_1^{(0)}, \eta_{21}^{(0)}, v_{21}^{(0)}, \eta_3^{(0)}, v_3^{(0)}, \eta_{4j}^{(0)}, v_{4j}^{(0)}$, $j = 1, \dots, J$, by fitting the gamma priors on the posterior samples for $\tau_1, \tau_{21}, \tau_3$ and τ_{4j} , $j = 1, \dots, J$. These values are shown in Table 3. Finally, with these proper priors we ran our algorithm to obtain posterior samples. Specifically, we obtained $M = 1,000$ iterates $(\mathbf{p}_i^{(h)}, \hat{\delta}_i^{(h)})$, $h = 1, \dots, M$, $i = 1, \dots, c$. Inference about the \mathbf{p}_i, δ_i and any function of them can be made using these iterates in a straightforward manner.

Table 3

Estimates of $\eta^{(0)}$ and $v^{(0)}$ corresponding to the gamma densities on τ_1, τ_{21} for 45+ and $\tau_3, \tau_{41}, \tau_{42}, \tau_{43}$ for 45- by race and sex

Race	Sex	Age					
		45-				45+	
		τ_3	τ_{41}	τ_{42}	τ_{43}	τ_1	τ_{21}
W	M	$\eta^{(0)}$	3.698	2.341	3.085	2.685	4.408
		$v^{(0)}$.036	.071	.201	.163	.009
	F	$\eta^{(0)}$	4.200	3.294	2.481	1.819	4.788
		$v^{(0)}$.030	.059	.072	.017	.008
B	M	$\eta^{(0)}$	4.948	2.922	3.156	2.404	5.971
		$v^{(0)}$.068	.096	.169	.147	.107
	F	$\eta^{(0)}$	3.745	3.084	1.893	2.350	3.292
		$v^{(0)}$.055	.036	.049	.116	.009

4. AN ANALYSIS OF THE NHANES III DATA

In this section we illustrate our methodology using the BMI data from NHANES III. First, we study our estimates based on summary measures over the counties. Specifically, we use the weighted posterior distributions of the p_{ij} ,

$$\tilde{p}_j = \sum_{i=1}^c n_i p_{ij} / \sum_{i=1}^c n_i, j = 1, 2, 3$$

and the weighted posterior distribution of the δ_i

$$\tilde{\delta} = \sum_{i=1}^c n_i \delta_i / \sum_{i=1}^c n_i$$

for each of the eight age-race-sex domains. Then, for the first four examples in Table 2 we show small area effects.

We also show how to relate the p_{ijk} and the π_{ij} to age, race and sex using linear and nonlinear logistic regression models

4.1 Data Analysis

First, we performed a sensitivity analysis to assess the specifications of $\eta^{(0)}$ and $v^{(0)}$. We compared three choices of hyper-parameters $\Omega = (\eta^{(0)}, v^{(0)})$ to check the sensitivity of the specification of the hyper-parameters on inference. Our first choice is 4 times of Ω , i.e., $4\Omega = (4\eta^{(0)}, 4v^{(0)})$; our second choice is the hyper-parameters without any change, i.e., $\Omega = (\eta^{(0)}, v^{(0)})$; and our third choice is one fourth of Ω i.e., $\Omega/4 = (\eta^{(0)}/4, v^{(0)}/4)$.

Table 4 shows the simulation results for the sensitivity to the inference of \tilde{p}_j for the younger group (45-). The point estimates and standard deviations of the proportions are very similar over the three choices of hyper-parameters. Similarly, Table 5 shows the simulation results for \tilde{p}_j for the older group (45+). The point estimates for males are very similar over the three choices of the hyper-parameters, but there are small changes in the point estimates for females from 4Ω to Ω . The standard deviations are increased when Ω decreases for the females, but no substantial changes are detected for males. Generally, the nonignorable nonresponse model performs better than the ignorable nonresponse model, as the nonignorable nonresponse model is not sensitive to choices of the hyper-parameters.

Table 4

Sensitivity of \tilde{p}_j for choice of $\eta_3^{(0)}, v_3^{(0)}, \eta_{4j}^{(0)}$ and $v_{4j}^{(0)}$, $j = 1, \dots, 4$ for the younger group (45-) for the three BMI levels

Race	Sex	\tilde{p}_1	std(\tilde{p}_1)	\tilde{p}_2	std(\tilde{p}_2)	\tilde{p}_3	std(\tilde{p}_3)
(a) 4Ω							
W	M	.428	.022	.216	.019	.356	.022
	F	.476	.025	.232	.020	.292	.024
B	M	.419	.020	.212	.016	.369	.020
	F	.434	.026	.185	.023	.381	.027
(b) Ω							
W	M	.427	.022	.211	.020	.362	.025
	F	.476	.026	.223	.024	.301	.031
B	M	.419	.020	.208	.017	.373	.022
	F	.435	.025	.178	.026	.387	.029
(c) Ω/4							
W	M	.427	.022	.210	.021	.364	.027
	F	.475	.026	.220	.026	.304	.034
B	M	.419	.020	.206	.018	.375	.024
	F	.435	.025	.177	.028	.388	.029

Note 1: $\Omega = (\eta_3^{(0)}, v_3^{(0)}, \eta_{41}^{(0)}, v_{41}^{(0)}, \eta_{42}^{(0)}, v_{42}^{(0)}, \eta_{43}^{(0)}, v_{43}^{(0)})$.
Note 2: The nonignorable nonresponse model is applied to the younger group.

Table 5

Sensitivity of \tilde{p}_j for choice of $\eta_1^{(0)}, v_1^{(0)}, \eta_{21}^{(0)}, v_{21}^{(0)}$ for the older group (45+) for the three BMI levels

Race	Sex	\tilde{p}_1	std(\tilde{p}_1)	\tilde{p}_2	std(\tilde{p}_2)	\tilde{p}_3	std(\tilde{p}_3)
(a) 4Ω							
W	M	.030	.005	.306	.018	.664	.018
	F	.081	.002	.436	.004	.483	.004
B	M	.053	.011	.317	.017	.630	.018
	F	.075	.005	.201	.004	.724	.006
(b) Ω							
W	M	.031	.005	.292	.016	.677	.016
	F	.063	.002	.443	.006	.494	.005
B	M	.053	.011	.316	.019	.631	.020
	F	.066	.012	.237	.018	.697	.019
(c) Ω/4							
W	M	.031	.005	.293	.018	.676	.019
	F	.073	.015	.359	.011	.568	.019
B	M	.053	.010	.317	.018	.630	.019
	F	.065	.013	.221	.022	.714	.025

Note 1: $\Omega = (\eta_1^{(0)}, v_1^{(0)}, \eta_{21}^{(0)}, v_{21}^{(0)})$.
Note 2: The ignorable nonresponse model is applied to the older group.

Table 6

Point estimates and 95% credible intervals for the weighted probability of response, $\tilde{\delta} = \sum_{i=1}^c n_i \delta_i / \sum_{i=1}^c n_i$, for three choices of Ω and the younger group

		4 Ω			Ω			$\Omega/4$		
Race	Sex	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval	$\tilde{\delta}$	std($\tilde{\delta}$)	Interval
W	M	.775	.016	(.744, .805)	.769	.017	(.735, .801)	.767	.018	(.732, .799)
	F	.855	.017	(.821, .886)	.855	.020	(.810, .887)	.853	.022	(.806, .887)
B	M	.786	.016	(.752, .817)	.780	.018	(.740, .813)	.778	.018	(.739, .811)
	F	.880	.013	(.854, .902)	.878	.015	(.845, .903)	.876	.015	(.838, .903)

Note: See the note to Table 1.

Table 6 shows point estimates of the probability of responding $\tilde{\delta}$, and their 95% credible intervals for three choices of Ω . The probabilities of responding for males are lower than those for females, and this pattern remains the same for three choices of Ω . If a similar survey is conducted in the future, we should increase the sample size by 1.30 = (1/.769) times for white males and 1.17 = (1/.855) times for white females (e.g., if complete data are required from 1,000 households, the interviewer needs to contact 1,300 white males).

In Table 7 we present 95% credible intervals for the \tilde{p}_j for the three BMI levels. For the younger group, \tilde{p}_1 of BMI level 1 is the highest, and \tilde{p}_2 of BMI level 2 is the lowest. The lower bounds for \tilde{p}_1 and \tilde{p}_3 are similar for the younger group except for white females, and those for \tilde{p}_2 are similar except for the non-white females. For the older group, \tilde{p}_3 of BMI level 3 is highest, and \tilde{p}_1 of BMI level 1 is lowest. Specifically \tilde{p}_1 , \tilde{p}_2 are high and \tilde{p}_3 is low for the white males.

Table 7

95% credible intervals for the weighted proportions, $\tilde{p}_j = \sum_{i=1}^c n_i p_{ij} / \sum_{i=1}^c n_i$ by age, race and sex

		95% credible interval		
Age	Race Sex	\tilde{p}_1	\tilde{p}_2	\tilde{p}_3
45-	W	M (.382, .470)	(.174, .252)	(.314, .412)
		F (.425, .525)	(.171, .269)	(.243, .371)
	B	M (.381, .455)	(.176, .241)	(.333, .419)
45+	W	F (.385, .482)	(.130, .230)	(.329, .442)
		M (.022, .041)	(.255, .326)	(.643, .710)
	B	F (.059, .068)	(.431, .451)	(.486, .505)
		M (.035, .076)	(.282, .352)	(.592, .670)
		F (.040, .093)	(.206, .265)	(.661, .731)

Note 1: The nonignorable nonresponse model is applied to the younger group.

Note 2: The ignorable nonresponse model is applied to the older group.

As suggested by a referee, we have looked at the results for older white females (45+) in Table 7 in greater detail. From Table 1 the observed proportions in the three BMI levels are .079, .347 and .568. However, the 95% credible intervals for the population proportions in Table 7 are (.059, .068), (.431, .451) and (.486, .505) respectively. That

is, while the observed proportions are close to the intervals, none of these intervals contains the observed proportions. We can explain this phenomenon in the following manner. The data for older white females (45+) are very sparse. For the 34 counties the quartiles of the observed counts in the three BMI levels are (0,1,3), (3,6,10) and (5,9,14) respectively. Thus, when the ignorable nonresponse model is fit to the 34 counties, there is shrinkage not only across the counties but also across the BMI levels. Consequently, the largest proportion tends to be smaller and the smallest proportion tends to be larger, and since the three proportions must add up to one, the second proportion must also "shrink" somewhat. In addition, consider the sensitivity analysis in Table 5. We can approximate 95% credible intervals for \tilde{p}_1 , \tilde{p}_2 and \tilde{p}_3 , by using the posterior mean $\pm 2 \times$ standard deviation. The intervals at 4 Ω and Ω do not contain the observed proportions, but the intervals at $\Omega/4$ do. Therefore, because of the sparseness of the data, there is some sensitivity to inference for older white females (45+) with respect to the prior misspecification of Ω . These results are expected within the small area context, when there are sparse data.

We use the first four examples in Table 2 to illustrate small area estimation. As it can be imagined, it is too cumbersome to present all the estimates for the 34 counties and the 8 domains. Table 8 shows the posterior means, standard deviations and 95% credible intervals for the p_{ij} and the δ_i .

First, we compare the estimates of the p_{ij} from the ignorable and nonignorable nonresponse models. The estimates from the two models are generally different with the intervals for the nonignorable nonresponse model wider than those for the ignorable nonresponse model.

Second, we consider the estimates (based on the nonignorable nonresponse model) of p_{ij} for the individual counties in Table 8 with the overall averages, the \tilde{p}_j in Table 7. As expected, when the \tilde{p}_j are obtained, there is an overall reduction in variability because of the extra smoothing, thereby making the intervals for the smaller domains relatively much wider. In fact, all the intervals for the small domains contain the intervals for \tilde{p}_j .

Finally, in Table 8 we consider the estimates of \tilde{p}_{ij} for the individual counties with the overall average, \tilde{p}_j in Table 7. The message is similar to that for the p_{ij} .

However, we note that the first example is an exception where the credible interval for $\delta_i(.459, .773)$ is almost completely to the left side of the credible interval for $\tilde{\delta}(.735, .801)$. Thus, there is much shrinkage for this example which is due to the relatively large number of nonrespondents, 14 in this county for white males 45-.

Table 8

Comparison of the ignorable (ig) and the nonignorable (nig) nonresponse models for the four examples (Ex) corresponding to small domains using the cell probabilities (p_j) and the probability of responding (δ)

Ex	Model		p_1	p_2	p_3	δ
1	ig	avg	.444	.308	.248	
		std	.073	.067	.067	
		CI	(.297, .593)	(.193, .450)	(.125, .386)	
	nig	avg	.450	.276	.273	.637
		std	.093	.079	.082	.081
		CI	(.256, .638)	(.137, .444)	(.133, .448)	(.459, .773)
2	ig	avg	.480	.308	.213	
		std	.075	.066	.062	
		CI	(.324, .619)	(.193, .452)	(.097, .344)	
	nig	avg	.493	.263	.244	.879
		std	.074	.065	.062	.041
		CI	(.338, .628)	(.141, .406)	(.121, .394)	(.782, .948)
3	ig	avg	.420	.306	.274	
		std	.071	.063	.063	
		CI	(.276, .561)	(.192, .437)	(.161, .416)	
	nig	avg	.438	.252	.310	.741
		std	.079	.072	.074	.058
		CI	(.283, .591)	(.116, .406)	(.186, .483)	(.607, .836)
4	ig	avg	.448	.263	.288	
		std	.089	.075	.081	
		CI	(.278, .620)	(.127, .424)	(.138, .468)	
	nig	avg	.430	.261	.308	.874
		std	.100	.086	.091	.046
		CI	(.217, .619)	(.104, .453)	(.145, .517)	(.768, .948)

Note: For each parameter avg = posterior mean; std = posterior standard deviation; CI = 95% credible interval

4.2 Linear and Nonlinear Logistic Regression Models

Let q_{ijl} denote the probability that a respondent in l^{th} ($l = 1, 8$) age-race-sex group in the i^{th} county belongs to the j^{th} BMI level. (We add the subscript l to the p_{ij} to denote the domains.) Letting $v_{ijl} = \log \{ \sum_{\delta=1}^J q_{i\delta l} / (1 - \sum_{\delta=1}^J q_{i\delta l}) \}$, $j = 1, \dots, J-1$, we take

$$v_{ijl} = (\theta_j - (\mu_i + \alpha_l)) / \psi_i \quad (10)$$

subject to the constraints $\sum_{i=1}^c \mu_i = 0$, $\sum_{j=1}^{J-1} \theta_j = 0$, $\sum_{l=1}^8 \alpha_l = 0$, and $\sum_{i=1}^c \ln \psi_i = 0$. The parameters θ_j , μ_i , α_l and ψ_i in (10) have posterior distributions whose properties are inherited from the posterior distributions of q_{ijl} . Each iterate of the MH algorithm provides a value for q_{ijl} which is used in (10), and a nonlinear least squares problem is solved using an iterative method to get the values of θ_j , μ_i , α_l and ψ_i (see Appendix 2). Alternatively, we can

also use the much simpler linear logistic model in which the ψ_i in (10) are taken equal to unity. In this case, the least squares estimators of θ_j , μ_i , α_l and α_l exist in closed form at the h^{th} iteration of MH algorithm. Specifically, for $\phi_i = 0$, we have the least squares estimates $\hat{\mu}_i = \bar{v} \dots - \bar{v}_{i.}$, $\hat{\theta}_j = \bar{v}_j - \bar{v}_{..}$, $\hat{\alpha}_l = \bar{v} \dots - \bar{v}_{..l}$, where

$$\bar{v} \dots = \sum_{i=1}^c \sum_{j=1}^{J-1} \sum_{l=1}^8 v_{ijl} / 8c (J-1),$$

$$\bar{v}_{i.} = \sum_{j=1}^{J-1} \sum_{l=1}^8 v_{ijl} / 8 (J-1),$$

$$\bar{v}_{..l} = \sum_{i=1}^c \sum_{j=1}^{J-1} v_{ijl} / 8c$$

and $\bar{v}_{..l} = \sum_{i=1}^c \sum_{j=1}^{J-1} v_{ijl} / c (J-1)$. The nonlinear least squares problem is solved using an iterative method to get the values of θ_j , μ_i and α_l .

We present 95% credible intervals for θ_1 , θ_2 and $\alpha_1, \dots, \alpha_8$ for the younger and older groups by regression type in Table 9. For the cut-points θ_j , θ_1 gives a large negative effect compared to θ_2 . The relative measure $\alpha_l (l = 1, \dots, 4)$ of the younger group gives a negative effect, while the relative measure $\alpha_l (l = 5, \dots, 8)$ of the older group gives positive effects. The 95% credible intervals for linear and nonlinear estimates are essentially the same.

We also relate the probability of response, $\delta_i = \sum_{j=1}^J \pi_{ij} p_{ij}$, to race and sex using linear and nonlinear logistic regression models for the younger group. The 95% credible intervals for θ and $\alpha_1, \dots, \alpha_4$ for the young group by regression type are shown in Table 10. Credible intervals for all α_l for the nonlinear model are shorter than those for the linear model. However, for the nonlinear model the credible interval for θ is wider than and on the right of that for the linear model.

Table 9

Comparison of 95% credible intervals for θ_1 , θ_2 and $\alpha_1, \dots, \alpha_8$ for both younger and older groups by regression type

	Linear	Nonlinear
θ_1	(-1.743, -1.469)	(-1.731, -1.466)
θ_2	(0.028, 0.196)	(0.025, 0.193)
α_1	(-1.167, -0.751)	(-1.159, -0.751)
α_2	(-1.395, -0.939)	(-1.385, -0.937)
α_3	(-1.127, -0.723)	(-1.119, -0.728)
α_4	(-1.112, -0.659)	(-1.103, -0.658)
α_5	(1.198, 1.514)	(1.188, 1.498)
α_6	(0.513, 0.689)	(0.506, 0.685)
α_7	(0.715, 1.210)	(0.725, 1.225)
α_8	(0.809, 1.310)	(0.803, 1.300)

Table 10
Comparison of 95% credible intervals for θ and $\alpha_1, \dots, \alpha_4$ for the younger group by regression type

	Linear	Nonlinear
θ	(1.455, 1.729)	(1.664, 2.174)
α_1	(0.165, 0.592)	(0.146, 0.523)
α_2	(-0.535, 0.014)	(-0.467, 0.007)
α_3	(0.078, 0.546)	(0.079, 0.484)
α_4	(-0.704, -0.165)	(-0.638, -0.169)

5. A SIMULATION STUDY

We describe a small simulation study to assess the performance of our multinomial nonignorable nonresponse model. We focus on the probability of responding.

We use the observed data from younger white males to obtain the posterior means of p_{i1}, p_{i2}, p_{i3} and $\pi_{i1}, \pi_{i2}, \pi_{i3}$ for each county. These are taken to be the true (t) values which we denote by $p_{i1}^{(t)}, p_{i2}^{(t)}, p_{i3}^{(t)}$ and $\pi_{i1}^{(t)}, \pi_{i2}^{(t)}, \pi_{i3}^{(t)}$. Thus, the true probability of responding in the i^{th} county is $\delta_i^{(t)} = \sum_{j=1}^3 p_{ij}^{(t)} \pi_{ij}^{(t)}$ and the weighted probability of responding is $\tilde{\delta}^{(t)} = \sum_{i=1}^c n_i \delta_i^{(t)} / \sum_{i=1}^c n_i$. In our simulated examples, we used the n_i as in the BMI data for younger white males, and we kept the $p_{ij}^{(t)}$ fixed throughout. However, we varied the π_{ij} in the following manner. We kept π_{i1} fixed at $\pi_{i1}^{(t)}$, and we denote the vector of the π_{i1} by π_1 . The 34 values of the π_1 range from .73 to .83. Then, we set $\pi_2 = a\pi_1$ and $\pi_3 = b\pi_1$, where $a, b = 0.8, 0.9, 1.0$. (We denote the vectors of the π_{i2} and the π_{i3} by π_2 and π_3 respectively.) Thus, there are 9 simulated examples.

Then, for each (a, b) we generated counts for a multinomial probability mass function with probabilities $p_{i1}^{(t)} \pi_{i1}, p_{i2}^{(t)} \pi_{i2}, p_{i3}^{(t)} \pi_{i3}, p_{i1}^{(t)} (1 - \pi_{i1}), p_{i2}^{(t)} (1 - \pi_{i2}), p_{i3}^{(t)} (1 - \pi_{i3})$. We denote these cell counts by $y_{i1}, y_{i2}, y_{i3}, z_{i1}, z_{i2}, z_{i3}$ and the number of respondents is $r_i = \sum_{j=1}^3 y_{ij}$. Then, we fit the nonignorable nonresponse model to the above data using the MH sampler, and we obtained $M = 1,000$ values $(p_{ij}^{(h)}, \pi_{ij}^{(h)}), h = 1, \dots, M$. For each value, we computed $\tilde{\delta}^{(h)} = \sum_{i=1}^c n_i \delta_i^{(h)} / \sum_{i=1}^c n_i$ where $\delta_i^{(h)} = \sum_{j=1}^3 p_{ij}^{(h)} \pi_{ij}^{(h)}$.

In Table 11 we report posterior means, standard deviations, numerical standard errors (using the batch means method) and 95% credible interval for the probability of responding for each choice of (a, b) . We also computed $\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} | y, r)$ by counting the number of $\tilde{\delta}^{(h)}$ that are as large as $\tilde{\delta}^{(t)}$. An extremely large or small value of this latter quantity suggests model failure.

We plotted the estimates of the posterior densities of $\tilde{\delta}$ by choices of a and b which we obtained by using normal kernel density estimator with an optimal window width from an output analysis of the MH algorithm. The densities are a unimodal, peaked and almost symmetric. By increasing (a, b) from (0.8, 0.8) to (1.0, 1.0), the mode of the posterior densities increase.

Table 11
Characteristics of the probability of responding

π_2	stat	π_3		
		$0.8 * \pi_1$	$0.9 * \pi_1$	$1.0 * \pi_1$
$0.8 * \pi_1$	true	0.690	0.719	0.748
	avg	0.712	0.739	0.764
	std	0.016	0.015	0.014
	nse	0.0030	0.0031	0.0029
	CI	(0.678, 0.742)	(0.708, 0.767)	(0.734, 0.750)
$0.9 * \pi_1$	prob	0.082	0.095	0.135
	true	0.706	0.735	0.764
	avg	0.710	0.742	0.776
	std	0.017	0.016	0.014
	nse	0.0030	0.0031	0.0031
$1.0 * \pi_1$	CI	(0.673, 0.742)	(0.712, 0.769)	(0.745, 0.802)
	prob	0.377	0.303	0.210
	true	0.722	0.751	0.780
	avg	0.726	0.758	0.784
	std	0.017	0.015	0.015
	nse	0.0036	0.0036	0.0026
	CI	(0.693, 0.757)	(0.725, 0.784)	(0.750, 0.809)
	prob	0.399	0.318	0.380

Note: avg = posterior mean; std = standard deviation; nse = numerical standard error; CI = 95% credible interval; prob = $\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} | y, r)$; the 34 values of π_1 range from .73 to .83.

In Table 11 we show that all the credible intervals contain the true values and the posterior means are close to the true value with the least discrepancy for the near ignorable nonresponse cases. The standard deviations are very similar across the nine simulated examples. Also, the numerical standard errors (nse) are small and similar for all nine simulated examples. The estimates of $\Pr(\tilde{\delta} < \tilde{\delta}^{(t)} | y, r)$ range from 0.30 to 0.40, except for the most nonignorable nonresponse cases in which $(a, b) = (.8, .8)$ and $(.8, .9)$. Thus, the model does perform reasonably well.

6. CONCLUSION

We have described a Bayesian methodology that can be used to analyze multinomial data for small areas when there is nonignorable nonresponse. A hierarchical model is used, and we have shown that it performs reasonably well. In fact, we have extended the method of Stasny (1991) in two directions: (a) we have considered multinomial data with more than two cells (binomial) and (b) we have done a full Bayesian analysis. Both (a) and (b) have been implemented for small areas.

The Markov chain Monte Carlo method permits an assessment of the complex structure of the multinomial nonresponse estimation. Our empirical analysis and simulation study indicate good performance of the model for these data. Thus, the method of ratio estimation currently

used in NHANES III may be replaced by our Bayesian method as the nonrespondents' characteristics might differ from those of the respondents. In fact, an application of our model to the NHANES III data shows that in each county there are substantial differences in the proportions of individuals at the three BMI levels by age and sex. This can be seen in Table 1 when the observed counts are summed over the counties. But, we have obtained inference (including measure of precision) for each county by age, race and sex.

Our methodology can be extended in three ways. First, it is feasible to use a model that incorporates an extent of nonignorability, rather than just the dichotomy of ignorable nonresponse and nonignorable nonresponse. Second, one can use other prior distributions (*e.g.*, Dirichlet process prior) to model heterogeneity in the clustering of the areas rather than assuming homogeneity of the areas as we have done. Third, one can use a fourth stage in our model to accommodate clustering within households as well as clustering within areas (counties) in NHANES III. These tasks are very difficult.

ACKNOWLEDGEMENT

This work was done at the National Center For Health Statistics while Balgobin Nandram was the first ASA/NCHS Research Fellow and Geunshik Han was on sabbatical leave from Hanshin University, Korea.

APPENDIX 1

Metropolis-Hastings Samplers

For the ignorable nonresponse model, (μ_1, τ_1) and (μ_{21}, τ_{21}) are independent a posteriori with

$$p(\mu_1, \tau_1 | \mathbf{y}, \mathbf{r}) \propto p(\mu_1, \tau_1) \prod_{i=1}^c \left\{ \frac{D(\mathbf{y}_i + n_i - r_i + \mu_1 \tau_1)}{D(\mu_1 \tau_1)} \right\} \quad (\text{A.1})$$

and

$$p(\mu_{21}, \tau_{21} | \mathbf{y}, \mathbf{r}) \propto p(\mu_{21}, \tau_{21})$$

$$\prod_{i=1}^c \left\{ \frac{B(r_i + \mu_{21} \tau_{21}, r_i - y_i + (1 - \mu_{21}) \tau_{21})}{B(\mu_{21} \tau_{21}, (1 - \mu_{21}) \tau_{21})} \right\} \quad (\text{A.2})$$

where $p(\mu_1, \tau_1)$ and $p(\mu_{21}, \tau_{21})$ are the prior distributions. Samples can be obtained from each of (A.1) and (A.2) using the MH algorithm of Nandram (1998).

For the nonignorable nonresponse model, it is convenient to condition on \mathbf{z} to obtain

$$p(\mu_3, \tau_3 | \mathbf{z}, \mathbf{y}, \mathbf{r}) \propto p(\mu_3, \tau_3) \prod_{i=1}^c \left\{ \frac{D(\mathbf{y}_i + \mathbf{z}_i + \mu_3 \tau_3)}{D(\mu_3 \tau_3)} \right\} \quad (\text{A.3})$$

$$p(\mu_{4j}, \tau_{4j} | \mathbf{z}, \mathbf{y}, \mathbf{r}) \propto p(\mu_{4j}, \tau_{4j})$$

$$\prod_{i=1}^c \left\{ \frac{B(y_{ij} + \mu_{4j} \tau_{4j}, z_{ij} + (1 - \mu_{4j}) \tau_{4j})}{B(\mu_{4j} \tau_{4j}, (1 - \mu_{4j}) \tau_{4j})} \right\}, \quad (\text{A.4})$$

where $p(\mu_3, \tau_3), p(\mu_{4j}, \tau_{4j}), j = 1, \dots, J$ are the prior distributions. Given \mathbf{z} , (A.3) and (A.4) are independent with

$$p(z_{i1} = t_{i1}, \dots, z_{iJ} = t_{iJ} | \mathbf{y}, \mathbf{r}, \mu_4, \tau_4, \mu_{3j}, \tau_{3j}, j = 1, \dots, J) = w_{i_{i1}t_{i2} \dots t_{iJ}} / \sum_{t_{i1}=0}^{n_i-r_i} \dots \sum_{t_{iJ}=0}^{n_i-r_i} w_{i_{i1}t_{i2} \dots t_{iJ}}, \quad (\text{A.5})$$

for $t_{ij} = 0, 1, \dots, n_i - r_i, \sum_{j=1}^J t_{ij} = n_i - r_i$,

$$w_{i_{i1}t_{i2} \dots t_{iJ}} = \binom{n_i - r_i}{t_{i1}, \dots, t_{iJ}} D(\mathbf{y}_i + \mathbf{t}_i + \mu_3 \tau_3) \prod_{j=1}^J B(y_{ij} + \mu_{4j} \tau_{4j}, t_{ij} + (1 - \mu_{4j}) \tau_{4j}).$$

We ran the MH sampler by drawing a random deviate from each of (A.3), (A.4), and (A.5). It is easy to draw a random deviate from (A.5). Samples were obtained from each of (A.3), (A.4) and (A.5) using the MH algorithm of Nandram (1998).

APPENDIX 2

Nonlinear least squares estimates

Let

$$v_{ijl} = \log \left\{ \sum_{s=1}^j q_{isl} / \left(1 - \sum_{s=1}^j q_{isl} \right) \right\}, j = 1, \dots, J-1 = J'.$$

These v_{ijl} are obtained for each iterate from the Metropolis-Hastings sampler. To solve the nonlinear least squares problem we minimized

$$\sum_{i=1}^c \sum_{j=1}^{J'} \sum_{l=1}^8 \left\{ v_{ijl} - e^{\eta_i} (\theta_j - (\mu_i + \alpha_l)) \right\}^2 \quad (\text{A.1})$$

subject to the constraints, $\sum_{i=1}^c \mu_i = 0, \sum_{j=1}^{J'} \theta_j = 0, \sum_{l=1}^8 \alpha_l = 0$, and letting $e^{\eta_i} = \psi_i, \sum_{i=1}^c \ln \psi_i = 0$.

Taking partial derivatives to find the least squares estimate, we have

$$\hat{\varphi}_i = \log \left\{ \frac{\sum_{j=1}^{J'} \sum_{l=1}^8 v_{ijl} (\hat{\theta}_j - \hat{\mu}_i - \hat{\alpha}_l)}{\sum_{j=1}^{J'} \sum_{l=1}^8 (\hat{\theta}_j - \hat{\mu}_i - \hat{\alpha}_l)^2} \right\} = \log \psi_i^{-1} \quad (\text{A.2})$$

where

$$\hat{\theta}_j = \left[\sum_{i=1}^c e^{2\hat{\phi}_i} \left\{ \frac{1}{8} \sum_{t=1}^8 \left(e^{-\hat{\phi}_i} v_{ijt} + \hat{\mu}_i + \hat{\alpha}_t \right) \right\} \right] / \sum_{i=1}^c e^{2\hat{\phi}_i}, \quad (\text{A.3})$$

$$\hat{\mu}_i = \left(\frac{1}{8J'} \right) \sum_{t=1}^8 \sum_{j=1}^{J'} \left\{ \hat{\theta}_j - \left(\hat{\alpha}_t + e^{-\hat{\phi}_i} v_{ijt} \right) \right\} \quad (\text{A.4})$$

and

$$\hat{\alpha}_t = \sum_{i=1}^c \frac{1}{J'} \sum_{j=1}^{J'} e^{2\hat{\phi}_i} \left\{ \hat{\theta}_j - \left(\hat{\mu}_i + e^{-\hat{\phi}_i} v_{ijt} \right) \right\} / \sum_{i=1}^c e^{2\hat{\phi}_i}. \quad (\text{A.5})$$

With these settings we draw the q_{ijt} from a MH algorithm, and the nonlinear least squares problem is solved using an iterative method to get values of ϕ_j , θ_j , μ_i and α_t . Let

$$v_{ijt}^{(h)} = \log \left\{ \frac{\sum_{s=1}^j q_{ist}^{(h)}}{\left(1 - \sum_{s=1}^j q_{ist}^{(h)} \right)} \right\},$$

where $q_{ist}^{(h)}$ denotes the value of q_{ist} at the h^{th} iterate of the MH algorithm. Then we minimize (A.1) subject to the above constraints at the h^{th} iterate to obtain $\phi_i^{(h)}$, $\theta_j^{(h)}$, $\mu_i^{(h)}$ and $\alpha_t^{(h)}$. These iterates provide an estimate of the posterior distributions of ϕ_j , θ_j , μ_i and α_t . Convergence occurred for our application in less than 10 iterations.

REFERENCES

- ALBERT, J.H., and GUPTA, A.K. (1985). Bayesian methods for binomial data with applications to a nonresponse problem. *Journal of the American statistical Association*. 80, 167-174.
- BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American statistical Association*. 83, 62-69.
- BASU, D., and PEREIRA, C.A. (1982). On the Bayesian analysis of categorical data: The Problem of nonresponse. *Journal of Statistical Planning and Inference*. 6, 345-362.
- CRAWFORD, S. L., JOHNSON, W.G. and LAIRD, N.M. (1993). Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey (with discussions). In *Case Studies in Bayesian Statistics* (Eds. C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Sinpurwalla). New York: Springer-Verlag, 78-117.
- DE HEER, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*. 15, 129-142.
- DEELY, J.J., and LINDLEY, D.V. (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association*. 76, 833-841.
- FORSTER, J.J., and SMITH, W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical society, Series B*. 60, 57-70.
- GROVES, R.M., and COUPER, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*. 5, 475-492.
- KUCZMARSKI, R.J., CARROL, M.D., FLEGAL, K.M. and TROIANO, R. P. (1997). Varying body mass index cutoff points to describe overweight prevalence among U. S. adults: NHANES III (1988 to 1994). *Obesity Research*. 5, 542-548.
- KAUFMAN, G.M., and KING, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association*. 68, 670-678.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- MALEC, D., DAVIS, W. and CAO, X. (1999). Model-based small area estimates of over-weight prevalence using sample selection adjustment. *Statistics in Medicine*. 18, 3189-3200.
- MOHADJER, L., BELL, B. and WAKSBERG, J. (1994). National Health and Nutrition Examination Survey III-accounting for item nonresponse bias. *National Center for Health Statistics*.
- NANDRAM, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*. 61, 97-126.
- NATIONAL CENTER FOR HEALTH STATISTICS (1992). Third National Health and Nutrition Examination Survey. *Vital and Health Statistics Series*. 2, 113.
- NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey. *Vital and Health Statistics, Series* 1, 32.
- OLSON, R.L. (1980). A least squares correction for selectivity bias. *Econometrica*. 48, 1815-1820.
- PARK, T. (1998). An approach to categorical data nonignorable nonresponse. *Biometrics*. 54, 1579-1590.
- PARK, T., and BROWN, M.B. (1994). Models for categorical data with nonignorable non-response. *Journal of the American Statistical Association*. 89, 44-52.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*. 63, 581-590.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- SCHAFER, J.L., EZZATI-RICE, T.M., JOHNSON, W., KHARE, M., LITTLE, R.J.A. and RUBIN, D.B. (1996). The NHANES III multiple imputation project. *Survey Research Methods, Proceedings of the American Statistical Association*. 28-37.
- STASNY, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association*. 86, 296-303.
- STASNY, E.A., KADANE, J.B. and FRITSCH, K.S. (1998). On the fairness of death penalty jurors: A comparison of Bayesian models with different levels of hierarchy and various missing-data mechanisms. *Journal of the American Statistical Association*. 93, 464-477.

Assessing the Bias Associated with Alternative Contact Strategies in Telephone Time-Use Surveys

JAY STEWART¹

ABSTRACT

In most telephone time-use surveys, respondents are called on one day and asked to report on their activities during the previous day. Given that most respondents are not available on their initial calling day, this feature of telephone time-use surveys introduces the possibility that the probability of interviewing the respondent about a given reference day is correlated with the activities on that reference day. Furthermore, noncontact bias is a more important consideration for time-use surveys than for other surveys, because time-use surveys cannot accept proxy responses. Therefore, it is essential that telephone time-use surveys have a strategy for making subsequent attempts to contact respondents. A contact strategy specifies the contact schedule and the field period. Previous literature has identified two schedules for making subsequent attempts: a convenient-day schedule and a designated-day schedule. Most of these articles recommend the designated-day schedule, but there is little evidence to support this viewpoint. In this paper, we use computer simulations to examine the bias associated with the convenient-day schedule and three variations of the designated-day schedule. The results support using a designated-day schedule, and validate the recommendations of the previous literature. The convenient-day schedule introduces systematic bias: time spent in activities done away from home tends to be overestimated. More importantly, estimates generated using the convenient-day schedule are sensitive to the variance of the contact probability. In contrast a designated-day-with-postponement schedule generates very little bias, and is robust to a wide range of assumptions about the pattern of activities across days of the week.

KEY WORDS: Telephone time-use surveys; Contact strategies; Bias; Computer simulations.

1. INTRODUCTION

Telephone time-use surveys present a unique data collection challenge because respondents are called on one day and asked to report on their activities during the previous day. The challenge arises because most respondents – about 75% (Kalton 1985) – are not contacted on their original calling day, necessitating additional contact attempts. In most surveys, it does not matter when these additional attempts are made, because respondents are being asked to report about a fixed reference period. And in most surveys recall does not suffer too much if respondents are contacted several days after the initial calling day. But in time-use surveys, respondents' ability to recall their activities on a given day falls off dramatically after a day or so, which means that the respondent must be assigned a new reference day if no contact is made on the initial calling day. As we will see below, this scenario introduces the possibility that the probability of interviewing the respondent about a given reference day is correlated with the activities on that reference day. Therefore it is essential that these surveys have a strategy for making subsequent attempts to contact respondents that does not introduce bias.

Contact Strategies

A contact strategy is comprised of a contact schedule and a field period. The contact schedule specifies which days of the week that contact attempts will be made, and the field period specifies the maximum number of weeks attempts will be made.

Contact schedules fall into two main categories: designated-day schedules and convenient-day schedules. Both types of schedule randomly assign each respondent to an initial calling day. If the respondent is contacted on the initial calling day, the interviewer attempts to collect information about the reference day, which is the day before the calling day. It is for subsequent contact attempts that these schedules differ.

Under a designated-day schedule, there are two approaches to making subsequent contact attempts. The interviewer could call the respondent on a later date, and ask the respondent to report activities for the original reference day. This approach maintains the original reference day, but extends the recall period. Harvey (1993) recommends allowing a recall period of no more than two days. The second approach is to postpone the interview and assign the respondent to a new reference day. Kalton (1985) recommends postponing the interview by exactly one week, so that the new reference day is the same day of the week as the original reference day.

These approaches are not mutually exclusive. For example, Statistics Canada's designated-day schedule allows interviewers to call respondents up to two days after the reference day (Statistics Canada 1999), and to postpone the interview by one week if the respondent cannot be reached after the second day of attempts. The interview can be postponed no more than three times (Statistics Canada). To illustrate, if the initial reference day is Monday the 1st, the respondent is called on Tuesday the 2nd and, if

¹ Jay Stewart, Office of Employment Research and Program Development, Bureau of Labor Statistics, 2 Massachusetts Avenue, NE Room 4945.

necessary, on Wednesday the 3rd. If no interview is obtained on either of these days, the respondent is called on Tuesday the 9th and, if necessary, on Wednesday the 10th, and asked to report on activities done on Monday the 8th. This process continues until the respondent is interviewed, refuses, or until four weeks pass.

The convenient-day schedule does not maintain the designated reference day. If no contact is made, the interviewer calls on the next day and each subsequent day until the respondent is contacted. Once contact is made, the interviewer attempts to complete the interview or, if the respondent is unwilling to complete the interview at that time, reschedule it to a day that is convenient for the respondent. The reference day is always the day prior to the interview. It is worth noting that because respondents are not likely to schedule interviews on busy days, allowing them to choose their interview day is really no different than the interviewer proposing consecutive days (or calling on consecutive days) until the respondent accepts. Hence, one may think of the convenient-day schedule as being functionally identical to an every-day contact attempt schedule.

A variant of the convenient-day schedule described above was used in the 1992-1994 Environmental Protection Agency (EPA) Time Diary Study conducted by the University of Maryland (see Triplett 1995). Respondents were not assigned to an initial calling day. Instead, they were assigned to either the weekday or the weekend sample. For example, those who were assigned to the weekend sample could be called on Sunday (to report about Saturday) or Monday (to report about Sunday). Interviewers were instructed to make at least 20 call attempts before finalizing the case as noncompleted.

Most methodological papers argue in favor of using a designated-day schedule (Kinsley and O'Donnell 1983; Kalton 1985; Lyberg 1989; Harvey 1993; and Harvey 1999). For example, Lyberg (1989) argues that the convenient-day schedule may introduce bias because "the respondent may choose a day when he/she is not busy, a day he/she is not engaged in socially unacceptable behavior, a day he/she thinks is representative, *etc.*" Kinsley and O'Donnell (1983) argue that the convenient-day schedule could exaggerate the number of events taking place outside the home, because the respondent is more likely to be interviewed on a day that immediately follows a day that he or she was out of the house.

Two of these studies directly compare the designated-day and convenient-day schedules (Kinsley and O'Donnell 1983; Lyberg 1989). In Kinsley and O'Donnell (1983), the experimental design divided the sample into two groups. They found that the two schedules produced similar response rates, and that the demographic composition was similar for both samples. They also found that the estimated time spent away from home was much higher under the convenient-day schedule than under the designated-day schedule. But it is impossible to determine whether the

convenient-day schedule overestimates time spent away from home or if the designated-day schedule underestimates time spent away from home, because the truth is not known. In Lyberg (1989), two diaries were collected from each respondent. One was collected using a designated-day schedule and the other was collected using a convenient-day schedule. However, the convenient-day diaries were conducted by an interviewer, while the designated-day diaries were self-administered several days after the convenient-day interview. So it is impossible to determine whether any differences were due to differences in contact schedules or whether they were due to mode effects.

Two studies (Lyberg 1989; Laaksonen and Pääkkönen 1992) investigate the effect of postponement on response rates. Both studies found that postponement increases response rates. Laaksonen and Pääkkönen (1992) also found that it was difficult to evaluate whether postponement introduces bias. Their results showed that respondents who postponed their interview spent less time on housekeeping and maintenance, and more time on shopping and errands. However, it is unclear whether these differences are the result of bias introduced by postponement, unobserved heterogeneity that is correlated with the postponement probability, or simply random noise. In any case, they argued that the differences were small, so that any bias was small.

One advantage of the convenient-day schedule is that it is possible to make many contact attempts in a short period of time. In contrast, the designated-day schedule – as proposed – permits only one contact attempt per week. So it is natural to ask: Would it be reasonable to modify the designated-day schedule to allow some form of day-of-week substitution? For example, if the respondent cannot be reached on Tuesday to report about Monday, would it be acceptable to contact the respondent on, say, Thursday and ask him or her to report about Wednesday? This modified schedule would allow for more contact attempts without having to extend the field period.

Because this type of substitution makes sense only if the substitute days are fairly similar to the original days, the first step was to determine which days, if any, were similar to one another. In earlier work, Stewart (2000) showed that Monday through Thursday are very similar to each other, Fridays are slightly different from the other weekdays, and Saturday and Sunday are very different from the weekdays and from each other. Hence, it would be reasonable to allow day-of-week substitution at least for Monday through Thursday.

Activity Bias and Noncontact Bias

When selecting a contact strategy, we need to be concerned with two types of bias: activity bias and non-contact bias. Activity bias occurs when the probability of contacting and interviewing a potential respondent on a particular day is correlated with his or her activities on that

day. Note that here and throughout the paper, the term contact probability refers to the probability of a productive contact (one that results in an interview). In order to isolate the effects of using alternative contact strategies, it is assumed that respondents always agree to an interview when contacted. Noncontact bias occurs when differences in contact probabilities across individuals are caused by differences in activities across individuals. Two simple numerical examples will illustrate these biases.

Example 1 – Activity Bias: Suppose that potential respondents' days fall into two categories: hard-to-contact (HTC) days and easy-to-contact (ETC) days. Further suppose that interviewers never contact respondents on HTC days (*i.e.*, that $P_H = 0$, where P_H is the contact probability on an HTC day), and that they always contact respondents on ETC days (*i.e.*, that $P_E = 1$, where P_E is the contact probability on an ETC day). Finally, suppose that the probability that any day is an ETC day is 0.5, so that on average half of each potential respondent's days are ETC and half are HTC. Note that all potential respondents are identical in the sense that the probability that any given day is an ETC day is 0.5 for all potential respondents. For simplicity, I assume that the activities of a given day can be summarized by an "activity index," I_J , where $I_J = 1 - P_J$ ($J = H, E$). The activity index represents time spent in activities that are negatively correlated with the contact probability. Thus, HTC days are days in which more time is spent in activities that are done away from home (working, shopping, active leisure, *etc.*), while ETC days are days in which more time is spent in activities that are done at home (housework, passive leisure, *etc.*). The true average activity index for the population of potential respondents is 0.5 ($= 0.5 \times 1 + 0.5 \times 0$).

If a convenient-day contact schedule is used and there is no limit on the number of call-backs, then HTC days are oversampled. To see why this occurs, it is instructive to work through the two possible contact sequences. If the initial contact attempt occurs on an ETC day, then the respondent is contacted and asked about the previous day (the diary day). Because HTC and ETC days are equally likely, on average half of these diary days will be HTC and the other half will be ETC. Therefore, the average activity index for the diary days of these respondents is equal to 0.5, which is the same as the population average. If, on the other hand, the initial contact day is an HTC day, then no interview takes place and the respondent is called on the following day. Contact attempts continue every day until the respondent is reached (on an ETC day). The average activity index for the diary days of these respondents is equal to one, because the respondent is always interviewed on an ETC day that immediately follows an HTC day. So if a given day is HTC (*i.e.*, the respondent does a lot of activities away from home), then it is more likely that that day will be selected as the reference day. Hence, the probability of interviewing the respondent on a given reference day is correlated with the activities on that

reference day. Since half of the initial contact attempts are made on HTC days and half are made on ETC days, the average activity index for the final sample is equal to 0.75 ($= 0.5 \times 0.5 + 0.5 \times 1$).

Example 2 – Noncontact Bias: Now suppose that potential respondents differ with respect to their contact probabilities, and that the contact probabilities for each individual do not vary from day to day. Suppose also that half of all potential respondents are HTC, with $P_H = 0.25$, and that the other half are ETC, with $P_E = 0.75$. If we attempt to contact each potential respondent four times, given these probabilities, virtually all (99.6%) ETC potential respondents are contacted. In contrast, only 68.4% of HTC potential respondents are contacted. The overall contact rate is 84% ($99.6\% \times 0.50 + 68.4\% \times 0.50$), but the final sample is not representative: 59.3% of the sample are ETC and only 40.7% are HTC. Therefore, estimates based on this sample will tend to underestimate the time spent in activities done by HTC people, and overestimate the time spent in activities done by ETC people.

The biases described above are not limited to time-use surveys. Although most surveys take steps to minimize noncontact bias, less attention has been devoted to activity bias. For example, in addition to their main focus on collecting event history information on employment, the National Longitudinal Surveys also include a few questions about labor force activities (employment and hours) during the week prior to the interview. Because these interviews tend to be scheduled at the convenience of the respondent, the respondent's activities during the reference week will be correlated with the probability of interviewing the respondent about that reference week. The intuition behind this correlation is exactly the same as that in Example 1. This correlation introduces bias into hours-worked estimates, although the direction of the bias is indeterminate. Hours worked per week tend to be overestimated for respondents who were unable to schedule an interview because of a heavy work schedule, and tend to be underestimated for respondents who were away on vacation. Activity bias is also an issue for travel surveys. Time spent away from home will tend to be overestimated if respondents are asked about, say, the four weeks prior to the interview. Asking respondents about a fixed reference period can eliminate this bias.

It is worth noting that noncontact bias is a more important consideration for time-use surveys than for other surveys, because, unlike most other surveys, time-use surveys cannot accept proxy responses. If proxy responses could be accepted then data on HTC individuals could be collected from proxies, who may be easier to contact. This would weaken the correlation between the individual's activities and the probability of collecting data about that individual.

The rest of the paper is organized as follows. In section 2, four contact strategies are introduced, and simple

simulations are used to assess the activity bias associated with each strategy. In section 3, the simulations are augmented with data from the May 1997 Work Schedule Supplement to the Current Population Survey and the 1992-94 University of Maryland Time Diary Study, and how the bias varies by specific activity is examined. In addition, the overall bias is decomposed to assess the relative contribution of activity bias and noncontact bias. Section 4 summarizes these results and makes recommendations.

2. CONTACT STRATEGIES, CORRELATED ACTIVITIES, AND ACTIVITY BIAS

In this section, the activity biases associated with the convenient-day schedule and each of the three variants of the designated-day schedule are compared. These schedules are defined as follows:

1. Convenient day (CD): Attempt to contact potential respondents every day following the initial contact attempt until contact is made or until the field period ends.
2. Designated day (DD): Attempt to contact potential respondents only once (no subsequent attempts).
3. Designated day with postponement (DDP): Attempt to contact potential respondents on the same day of the week as the initial attempt until contact is made or until the field period ends (as recommended by Kalton 1985).
4. Designated day with postponement and substitution (DDPS): Attempt to contact potential respondents every other day following the initial contact attempt until contact is made or until the field period ends.

The DDPS schedule assumes alternating Tuesday/Thursday and Wednesday/Friday contact days. Whether the first week is Tuesday/Thursday or Wednesday/Friday depends on the start day, which is randomly assigned.

As seen in Example 1, it is straightforward to show that a convenient-day schedule can introduce activity bias into time-use estimates when the base contact probability is the same each day (0.5) except for random noise (+0.5 with probability 1/2 or -0.5 with probability 1/2). Even though Stewart (2000) shows that Monday through Thursday are very similar on average, it is likely that the contact probabilities for some individuals vary systematically by day each week. For example, some individuals may be hard to contact on Monday, Wednesday, and Friday of each week. This systematic variation makes it considerably more complicated to determine whether sample estimates are biased, and to determine the direction and extent of that bias. One could model contact strategies and analytically solve for the bias under different assumptions about the pattern of contact probabilities. However, this is a cumbersome process, because each assumption about the pattern of

contact probabilities across days would require a separate solution. In contrast, computer simulations are an ideal way to assess the bias associated with alternative contact strategies under different assumptions about the pattern of contact probabilities. The computer program is simpler and produces more intuitive results than the analytical solution. And it is easy to modify the program to allow for different patterns. In section 3, realism is added to the simulations by incorporating real time-use data – something that would be impossible to do when taking an analytical approach.

Simulations

The simulation strategy was very straightforward. First, four weeks worth of “data” for each of 10,000 potential respondents was created. In order to focus on contact strategies, the sampling procedures are ignored and it is assumed that the sample of potential respondents is representative of the population. The simulations are designed to compare the four contact schedules above, so it is assumed that the “week” is five days long. Eligible diary days were restricted to Monday through Thursday, because, as noted above, these days are the most similar to each other. The next step was to simulate attempts to contact these respondents using the four contact schedules described above. Finally, the estimates generated using each schedule were compared to the true sample values.

To simplify the simulations I abstracted from specific activities, as in the examples above, and characterized each day using an activity index, I_j , ($J = H, E$) that ranges from 0 to 1. The activity index is given by $I_j = 1 - P_j$ where P_j is the probability of contacting and interviewing the respondent. To simulate the variation in activities across days, the contact probability on a given day is:

$$P_j = \bar{P}_j + \varepsilon,$$

where \bar{P}_j is the average contact probability on an HTC ($J = H$) or an ETC ($J = E$) day, and $\varepsilon \sim U(-\hat{\varepsilon}, \hat{\varepsilon})$. I assume that $\bar{P}_H < \bar{P}_E$, which means that, on average, respondents are less likely to be contacted on HTC days than on ETC days. To insure that contact probabilities lie in the $[0,1]$ interval, I set $\hat{\varepsilon}$ so that $\hat{\varepsilon} < \min(\bar{P}_H, 1 - \bar{P}_E)$.

There are many assumptions one can make regarding the pattern of activities across days. The simplest case is where all days are identical except for random noise. But as noted above, it is possible that potential respondents are systematically harder to contact on some days than others. To cover a wide range of activity patterns, the simulations were performed under the following eight assumptions about the pattern of HTC and ETC days in each of the four weeks:

1. Actual values of the activity index are distributed as $U(0,1)$, so that the average value is 0.5.
2. The first two days of every week are HTC and the last three days are ETC (HHEEE).
3. The first three days of every week are HTC and the last two days are ETC (HHHEE).

4. The first four days of every week are HTC and the last day is ETC (HHHHE).
5. The first day of every week is ETC and the last four are HTC (EHHHH).
6. The first two days of every week are ETC and the last three are HTC (EEHHH).
7. The first three days of every week are ETC and the last two are HTC (EEHH).
8. For half the sample Monday, Wednesday, and Friday are HTC and Tuesday and Thursday are ETC (HEHEH). For the other half of the sample the reverse is true (EHEHE).

In pattern 1, the base probability of contacting the respondent is the same, so that all of the variation in probabilities is due to the random term. In patterns 2-7, HTC days are grouped together either at the beginning of

the week or at the end of the week. And in pattern 8, the base probabilities alternate between HTC and ETC days. To focus on activity bias, separate simulations were performed for each of the 8 patterns described above. Thus, within a simulation all individuals have the same pattern of base probabilities.

Table 1 shows the results from a representative subset of the 153 simulations performed. The first four columns show the average contact probability on HTC and ETC days, the value of $\hat{\epsilon}$, and the true average activity index. The remaining columns contain estimates of the bias associated with the four contact schedules. The bias was computed as the difference between the estimated amount of time spent in each activity and the true amount of time spent in each activity, and then the difference was expressed as a percentage of the true value. Entries with an asterisk indicate that the bias is statistically different from the zero at the 5% level.

Table 1
Activity Bias Associated with Each Contact Strategy Under Alternative Assumptions About the Correlation of Activities Across Days

Average Contact Probability					Estimated Bias (Expressed as a percent of the true activity index)			
Activity Pattern	Hard-to-contact days	Easy-to-contact days	$\hat{\epsilon}$	True Average Activity Index	CD	DD	DDP	DDPS
Identical Base Probabilities								
	0.50		0.10	0.500	0.7*	-0.1	0.0	0.1
	0.50		0.30	0.500	5.3*	-0.3	0.1	0.2
	0.50		0.50	0.500	15.1*	-0.9	0.4	0.7
Grouped Base Probabilities								
HHEEE	0.75	0.25	0.05	0.500	0.7	-10.7*	-4.7*	-13.8*
	0.75	0.25	0.25	0.500	5.2*	-10.9*	-4.8*	-13.9*
	0.60	0.40	0.05	0.500	-0.1	-2.2*	-0.7*	-2.8*
	0.60	0.40	0.20	0.500	2.5*	-2.6*	-0.7*	-2.5*
HHHEE	0.75	0.25	0.05	0.625	-2.7*	-9.7*	-4.0*	-12.7*
	0.75	0.25	0.25	0.625	0.8	-10.3*	-4.1*	-12.8*
	0.60	0.40	0.05	0.550	-0.4*	-1.8*	-0.6*	-2.5*
	0.60	0.40	0.20	0.550	1.9*	-2.4*	-0.5	-2.2*
HHHHE	0.75	0.25	0.05	0.750	0.1	-0.1	0.1	0.0
	0.75	0.25	0.25	0.750	2.3*	-0.5	0.2	0.2
	0.60	0.40	0.05	0.600	0.1*	0.0	0.0	0.0
	0.60	0.40	0.20	0.600	1.9*	-0.3	0.2	0.2
EHHHH	0.75	0.25	0.05	0.625	1.7*	1.0	1.4*	0.7
	0.75	0.25	0.25	0.625	4.2*	-0.3	1.2*	0.7
	0.60	0.40	0.05	0.550	1.1*	0.3	0.5*	0.3
	0.60	0.40	0.20	0.550	2.9*	0.0	0.6*	0.4
EEHHH	0.75	0.25	0.05	0.500	-18.2*	-17.1*	-4.3*	-21.7*
	0.75	0.25	0.25	0.500	-15.9*	-17.9*	-4.5*	-20.9*
	0.60	0.40	0.05	0.500	-2.0*	-2.2*	-0.4	-2.6*
	0.60	0.40	0.20	0.500	-0.4	-2.4*	-0.3	-2.6*
EEEEH	0.75	0.25	0.05	0.375	-16.6*	-17.6*	-5.5*	-20.3*
	0.75	0.25	0.25	0.375	-11.4*	-17.6*	-5.6*	-19.6*
	0.60	0.40	0.05	0.450	-2.0*	-2.3*	-0.4	-2.5*
	0.60	0.40	0.20	0.450	0.0	-2.5*	-0.5	-2.5*
Alternating Base Probabilities								
HEHEH/EHEHE	0.75	0.25	0.05	0.500	31.5*	26.4*	9.6*	28.5*
	0.75	0.25	0.25	0.500	34.7*	26.5*	9.7*	29.4*
	0.60	0.40	0.05	0.500	5.6*	4.5*	1.3*	5.1*
	0.60	0.40	0.20	0.500	7.8*	4.3*	1.2*	5.1*

Note: Asterisks indicate that the estimated average activity index is statistically different from the true value at the 5% level.

Pattern 1 – Identical Base Probabilities with Random Noise

This pattern is essentially the same as in the numerical example above. The main result is that all of the contact schedules generate unbiased estimates for the average activity index, except the CD schedule. As expected, the CD schedule overestimates the average activity index. More importantly, when using the CD schedule, the estimated average activity index – and hence the bias when activities are uncorrelated across days – is positively correlated with the variance of ε . As the variance increases from 0.003 ($\hat{\varepsilon} = 0.1$) to 0.083 ($\hat{\varepsilon} = 0.5$), the bias increases from less than 1% to 15%. One can see the intuition behind this result by noting that a large negative realization of ε on a particular day makes it less likely that the respondent will be contacted on that day, and hence, more likely that that day will become the diary day. None of the other contact schedules are sensitive to the variance of ε .

Patterns 2-7 – Grouped Base Probabilities

The results are mixed when HTC days are grouped at either the beginning or the end of the week. In the simulations where $\bar{P}_E - \bar{P}_H$ is relatively small (0.2), all of the contact schedules perform reasonably well. The absolute value of the bias is less than 3% in all cases. However, when $\bar{P}_E - \bar{P}_H$ is relatively large (0.5), there are significant differences in the bias associated with each contact schedule. The DDP schedule performs the best overall. The bias exceeds 5% (in absolute value) only in pattern 7 (EEEEH), for which the bias is -5.5% . In contrast, when using the DD and DDPS schedules, the bias is in the 10 – 14% range in patterns 2 (HHEEE), 3 (HHHEE), and in the 16-20% range in patterns 6 (EEHHH), and 7 (EEEEH). The differences between the DD and DDPS schedules and the DDP schedule for these patterns are significant, both statistically and in practical terms. In patterns 4 (HHHHE) and 5 (EHHHH) the DDP schedule performs slightly worse than the DD and DDPS schedules, but the bias is so small (less than 1.5%) that the difference is of no practical significance. The CD schedule fares somewhat better than the DD and DDPS schedules. The bias is less than 5%, except in patterns 6 and 7 where the bias is in the 11 – 18% range. As in pattern 1 above, the estimated average activity index increases with the variance of ε under the CD schedule, but not under any of the other schedules. And as can be seen from Table 1, in patterns where the bias is negative (patterns 6 and 7), an increase in the variance of ε decreases the bias.

Pattern 8 – Alternating Base Probabilities

All of the contact schedules generate biased estimates, because ETC days are undersampled. As above, all of the schedules perform reasonably well when $\bar{P}_E - \bar{P}_H$ is relatively small. The bias is in the 5-8% range for all

schedules except DDP, for which the bias is about 1%. However, when $\bar{P}_E - \bar{P}_H$ is large, all of the contact schedules generate significant bias. The bias of about 10% for the DDP schedule is higher than for the other patterns but it is smaller than the 25-35% bias for the other schedules. Again, these differences are significant statistically, and they are significant in practical terms.

The reason that the DDPS schedule generates a large activity bias is that contact attempts are made on two HTC days and then on two ETC days (or the reverse). This pattern results in contacting respondents on a relatively large fraction of ETC days, and hence, diary days will be disproportionately HTC days. Not surprisingly, if the DDPS schedule is modified so the respondent is contacted on the same two days each week, there is virtually no bias.

It is clear from these simulations that the activity bias associated with each contact schedule depends on the pattern of activities across days, the contact probabilities on HTC and ETC days, and the variance of those probabilities. However, it is also clear that the DDP schedule outperforms the other schedules regardless of the pattern assumed. If each pattern is viewed as a different type of respondent, then the overall bias (which includes both activity and noncontact bias) depends on the relative frequency of each type in the population. Information on the incidence of each type would allow one to measure the overall bias, and, for each strategy, decompose the overall bias it into the portion due to activity bias, and the portion due to noncontact bias. This is investigated in the next section.

3. AUGMENTED SIMULATIONS

If one is willing to make some additional assumptions, it is possible to augment the simulations using data from other sources. The first assumption is that individuals' work schedules are a reasonable proxy for the patterns of HTC and ETC days, so that work days correspond to HTC days and nonwork days correspond to ETC days. The second assumption is that it is possible to replicate an individual's week by taking one day from each of five individuals.

Data from the May 1997 Work Schedule Supplement to the Current Population Survey (CPS) were used to obtain information about individuals' work schedules. Note that because of the need to know the prevalence of each type of schedule for the entire population, nonworkers were also included. Table 2 shows the patterns of work (W) days and nonwork (N) days from the May 1997 CPS. Approximately 88% of all individuals fall into two patterns. Forty-eight percent work all five weekdays, and 39% do not work any weekdays. Another 4% work four weekdays and have either Friday or Monday off. The remaining individuals do not exhibit any discernible pattern. To simplify the simulations, it was assumed that individuals either worked all 5 weekdays (workers) or that they did not work any weekdays (nonworkers).

Table 2
Distribution of Work Schedules

Activity Pattern					Percent	Cumulative Percent
M	Tu	W	Th	F		
-	-	-	-	-	39.40	39.40
W	W	W	W	W	48.11	87.51
W	W	W	W	-	2.63	90.14
-	W	W	W	W	1.63	91.77
W	W	W	-	-	0.81	92.58
W	W	-	-	-	0.26	92.84
-	-	-	W	W	0.37	93.21
-	-	W	W	W	0.68	93.89
W	-	W	-	W	0.49	94.38
-	W	-	W	-	0.25	94.63
-	-	-	-	W	0.51	95.14
W	-	-	-	-	0.25	95.39
W	W	-	W	W	0.73	96.12
W	-	-	-	W	0.36	96.48
W	-	-	W	W	0.70	97.18
Other patterns					2.82	100.00
Total					100.00	

Note: A "W" indicates a workday, and a "-" indicates a nonwork day. Author's tabulations from the May 1997 Work Schedule Supplement to the CPS. Observations were weighted using supplement weights. The sample size is 89,746 observations.

To generate information on individual activities, data from the 1992-94 EPA Time Diary Study, conducted by the University of Maryland were used. This dataset contains time-diaries for a sample of 7,408 adults (see Triplett 1995). Because each individual was interviewed only once, there is only one observation per person. The following repeated sampling method was used to construct 8 weeks worth of data for a sample of 18,974 "individuals." The diary data were divided into workdays and nonwork days. A diary day was considered a workday if the individual did any paid work during the day. Workdays were assigned to workers and nonwork days were assigned to nonworkers. Mondays were drawn from Monday observations, Tuesdays were drawn from Tuesday observations, *etc.* No observation was used more than once for a given individual, but the same observation could be used for more than one individual. The final sample proportions look fairly similar to the proportions from the CPS. Fifty-eight percent of individuals in the final sample were workers and 42% were nonworkers, which is reasonably close to the ratio of workers to nonworkers (1.38 vs. 1.23) in the CPS.

To compute the contact probabilities, it was necessary to make a third assumption. Following Pothoff, Manton, and Woodbury (1993), the contact probability was assumed to be equal to the number of minutes spent in activities done at home (excluding sleeping) divided by the time spent in all activities other than sleep. This process for generating contact probabilities has two important properties: (1) the contact probability for a given day is related to the activities

done on that day, and (2) one group of potential respondents (workers) has a lower average probability of a productive contact (0.36 vs. 0.72).

Tables 3a and 3b summarize the bias estimates from the augmented simulations. Table 3a shows the bias estimates assuming a 4-week field period, and Table 3b shows the same estimates assuming an 8-week field period. Each of the first four columns contains estimates of the bias associated with the four contact strategies. The entries for each strategy and each 1-digit activity include estimates of the activity bias for workers and nonworkers, and an estimate of the overall bias. The overall bias includes noncontact bias, so it is possible that the overall bias is larger (or smaller) than the activity bias for either group. The bias was computed as in the previous simulations, strategy and as before, an asterisk indicates that the bias is significantly different from the zero at the 5% level. The fifth column shows the true time spent in each activity by group and overall.

Comparing Tables 3a and 3b, we can see that the main difference is that, except for the DD strategy for which the field period is irrelevant, the overall bias is smaller when the field period is 8 weeks. This smaller overall bias is due mainly to the increased number of contact attempts, which disproportionately increases the probability that workers are contacted and makes the sample more representative (see Table 4). In contrast, estimates of the activity bias associated with the various contact strategies are not sensitive to the length of the contact period. The rest of this discussion will focus on the results in Table 3b.

The DD strategy generated virtually no activity bias. There were a few activities – Active Leisure, Entertainment/Socializing, Organizational Activities, Education/Training, and Active Child Care for workers, and Active Child Care for nonworkers – for which the activity bias was rather large, but none of these bias estimates are statistically significant. The overall bias for the DD strategy is quite large for most activities, which, as will be seen below, is primarily due to noncontact bias.

Comparing the other three strategies, one can see two patterns emerge. First, activity bias is significantly smaller (and generally not statistically significant) when using the DDP strategy or the DDPS strategy than when using the CD strategy. Second, the bias in the CD estimates follows the expected pattern. The bias tends to be positive for activities that are done away from home (Active Leisure, Entertainment/Socializing, Organizational Activities, Education/Training, Purchasing Goods/Services, and Paid Work), and negative for activities done at home (Passive Leisure, Personal Care, Active Child Care, and Housework). This pattern is consistent with research cited in the introduction that finds that reported time spent away from home is greater under a convenience-day strategy than under a designated-day strategy. More important, it is now clear that this finding is due to bias in convenient-day strategies rather than bias in designated-day strategies.

Table 3a
Estimated Bias – Augmented Simulations (4 Week Field Period)

Activity/Emp. Status Employment Status	CD	DD	DDP	DDPS	Time Spent in Activity (Truth)
Passive Leisure					
Nonworkers	-8.44*	0.12	-1.54	-1.03	314.72
Workers	-5.40*	1.07	0.43	0.82	152.04
Overall	-8.62*	13.56*	2.53*	0.38	220.70
Active Leisure					
Nonworkers	9.80*	-2.75	0.99	-0.66	65.94
Workers	-0.07	-7.34	-4.69	1.91	26.89
Overall	4.03*	11.75*	3.31	1.08	43.37
Entertainment/Socializing					
Nonworkers	19.41*	-2.01	-0.25	-1.20	67.30
Workers	8.63*	7.14	5.21	3.72	27.87
Overall	13.11*	15.78*	5.64*	1.37	44.51
Organizational Activities					
Nonworkers	19.58*	-0.98	9.00	3.84	19.25
Workers	13.77*	6.95	7.17	7.48	8.72
Overall	15.24*	15.26*	12.37*	5.99	13.16
Education/Training					
Nonworkers	32.77*	-0.42	12.54*	8.92*	43.60
Workers	-1.17	7.63	0.57	1.59	13.16
Overall	19.17*	22.02*	15.39*	8.00*	26.01
Personal Care					
Nonworkers	-0.50	-0.29	-0.49	-0.44	663.04
Workers	-0.52*	0.01	-0.06	-0.13	580.71
Overall	-0.79*	2.20*	0.34	-0.15	615.46
Purchasing Goods/Services					
Nonworkers	12.62*	1.35	0.11	-1.28	72.98
Workers	-4.05	4.62	-3.62	-5.43*	23.28
Overall	4.67*	22.36*	4.25*	-1.49	44.25
Active Child Care					
Nonworkers	-7.89*	5.11	-1.06	-0.54	24.13
Workers	-7.69*	-6.05	-4.09	-0.92	12.64
Overall	-9.09*	14.21*	0.77	-0.09	17.49
Housework					
Nonworkers	-8.88*	1.71	0.33	2.27	169.04
Workers	-10.55*	0.85	-2.03	-0.14	57.92
Overall	-11.49*	20.77*	4.53*	2.52*	104.82
Paid Work					
Nonworkers	—	—	—	—	—
Workers	2.95*	-0.77	0.25	-0.27	536.77
Overall	6.74*	31.44*	-7.74*	-1.87*	310.22

Note: Asterisks indicate that the bias in the estimated time spent in the activity is significantly different from zero at the 5% level.

Noncontact Bias

In general, the contact rate increases and the sample becomes more representative as the number of contact attempts increases (see Table 4). The contact rate is the lowest under the DD strategy (40%), and the sample is the least representative. Under both the DDP and the DDPS schedules, the contact rate increases and the sample becomes more representative as the field period increases from 4 to 8 weeks. Using a DDPS schedule with an 8-week field period (16 contact attempts) results in a contact rate of 80% and a representative sample. Not surprisingly, the sample generated by the DDP schedule with an 8 week field period is virtually identical to the one generated by the DDPS schedule with a 4 week field period.

Activity Bias vs. Noncontact Bias

To get a clearer picture of the contribution of each type of bias to the overall bias, the overall bias was decomposed into the portion due to activity bias, the portion due to noncontact bias, and the portion due to the interaction between the two biases. The overall bias for activity a and group g (workers or nonworkers) is given by:

$$F_g^* X_{ag} - F_g^* X_{ag}^* = F_g^* (X_{ag} - X_{ag}^*) + X_{ag}^* (F_g - F_g^*) + (F_g - F_g^*) (X_{ag} - X_{ag}^*)$$

Activity + Noncontact + Interaction

Table 3b
Estimated Bias – Augmented Simulations (8 Week Field Period)

Activity/Emp. Status	CD	DD	DDP	DDPS	Time Spent in Activity (Truth)
Employment Status					
Passive Leisure					
Nonworkers	-8.63*	-0.09	-1.62	-1.21	315.38
Workers	-5.24*	1.28	0.39	1.10	151.72
Overall	-8.72*	-13.51*	-0.35	-0.31	220.79
Active Leisure					
Nonworkers	10.62*	-2.03	1.76	0.06	65.46
Workers	0.00	-7.29	-3.50	2.21	26.87
Overall	4.49*	12.30*	0.50	0.82	43.16
Entertainment/Socializing					
Nonworkers	19.77*	-1.72	-0.15	-0.91	67.10
Workers	8.09*	6.64	5.52	2.76	28.00
Overall	13.06*	15.80*	2.47	0.40	44.50
Organizational Activities					
Nonworkers	18.92*	-1.53	8.59	3.25	19.36
Workers	14.03*	7.00	3.18	7.25	8.72
Overall	14.89*	14.88*	7.14*	4.76	13.21
Education/Training					
Nonworkers	33.56*	0.18	12.91*	9.55*	43.34
Workers	-0.72	8.24	0.77	2.01	13.09
Overall	19.73*	22.74*	10.29*	7.32*	25.86
Personal Care					
Nonworkers	-0.50	-0.29	-0.48	-0.44	663.03
Workers	-0.55*	0.00	-0.08	-0.16	580.81
Overall	-0.82*	2.20*	-0.17	-0.29	615.51
Purchasing Goods/Services					
Nonworkers	12.64*	1.36	-0.09	-1.28	72.97
Workers	-4.41	4.23	-3.66	-5.45*	23.36
Overall	4.48*	22.23*	-0.42	-2.58	44.30
Active Child Care					
Nonworkers	-7.67*	5.36	-1.04	-0.31	24.07
Workers	-8.02*	-6.18	-4.98	-1.65	12.66
Overall	-9.14*	14.30*	-2.23	-0.89	17.48
Housework					
Nonworkers	-9.02*	1.55	0.20	2.10	169.30
Workers	-10.55*	0.80	-2.15	-0.20	57.95
Overall	-11.64*	20.63*	0.17	1.34	104.94
Paid Work					
Nonworkers	—	—	—	—	—
Workers	2.96*	-0.78	0.30	-0.26	536.82
Overall	6.86*	-31.44*	-0.86	-0.22	310.25

Note: Asterisks indicate that the bias in the estimated time spent in the activity is significantly different from zero at the 5% level.

Table 4
Contact Rate Summary – Augmented Simulations

Field Period		CD	DD	DDP	DDPS	Truth
4 weeks	Contact Rate	89.68	40.35	71.79	78.39	
	Percent Nonworkers	40.08	60.07	46.82	43.14	42.21
	Percent Workers	59.92	39.93	53.18	56.86	57.79
8 weeks	Contact Rate	89.79	40.35	78.87	80.17	
	Percent Nonworkers	40.02	60.07	42.88	42.19	42.21
	Percent Workers	59.98	39.93	57.12	57.81	57.79

where F_g is the fraction of the sample in group g , and X_{ag} is the time spent in activity a by group g , and asterisks indicate the true values. The total bias for activity a is

obtained by summing this expression over workers and nonworkers, and is given by:

$$\sum_{g=W,N} (F_g X_{ag} - F_g^* X_{ag}^*) = \sum_{g=W,N} F_g^* (X_{ag} - X_{ag}^*) + \sum_{g=W,N} X_{ag}^* (F_g - F_g^*) + \sum_{g=W,N} (F_g - F_g^*) (X_{ag} - X_{ag}^*),$$

there are several things to take from these decompositions (shown in Table 5). First, under the CD schedule, all of the overall bias is due to activity bias. The large number of contact attempts virtually guarantees a representative sample, so that increasing the field period from 4 to 8 weeks

does not make much difference. In contrast, noncontact bias accounts for all of the bias under the DD schedule. Under both the DDP schedule and the DDPS schedule there is virtually no activity bias, and noncontact bias decreases dramatically as the field period is increased from 4 to 8 weeks. Not surprisingly, the noncontact bias for the DDP schedule with an 8-week field period is about the same as the noncontact bias under the DDPS schedule with a 4-week field period. In these simulations, the sample becomes fully representative when the field period is long enough to allow 16 contact attempts. Finally, the small magnitude of the interaction terms reflects the fact that activity and noncontact biases associated with each contact strategy are negatively correlated.

Table 5
Bias Decomposition – Augmented Simulations

	4 – week field period				8 – week field period			
	Total Bias	Activity Bias	Noncontact Bias	Interaction	Total Bias	Activity Bias	Noncontact Bias	Interaction
Passive Leisure								
CD	-8.62	-7.23	-1.57	0.18	-8.72	-7.29	-1.62	0.19
DD	13.56	0.50	13.16	-0.10	13.51	0.46	13.24	-0.18
DDP	2.53	-0.75	3.40	-0.11	-0.35	-0.83	0.50	-0.02
DDPS	0.38	-0.29	0.69	-0.02	-0.31	-0.30	-0.01	0.00
Active Leisure								
CD	4.03	6.27	-1.92	-0.32	4.49	6.80	-1.96	-0.35
DD	11.75	-4.40	16.08	0.06	12.30	-3.92	15.97	0.26
DDP	3.31	-1.05	4.15	0.20	0.50	-0.13	0.60	0.03
DDPS	1.08	0.26	0.84	-0.02	0.82	0.83	-0.02	0.00
Entertainment/Socializing								
CD	13.11	15.51	-1.89	-0.51	13.06	15.53	-1.92	-0.54
DD	15.78	1.30	15.82	-1.34	15.80	1.32	15.69	-1.21
DDP	5.64	1.72	4.08	-0.17	2.47	1.91	0.59	-0.02
DDPS	1.37	0.58	0.82	-0.04	0.40	0.42	-0.02	0.00
Organizational Activities								
CD	15.24	17.36	-1.70	-0.42	14.89	17.06	-1.76	-0.40
DD	15.26	2.05	14.28	-1.08	14.88	1.72	14.39	-1.23
DDP	12.37	8.30	3.69	0.39	7.14	6.53	0.54	0.07
DDPS	5.99	5.24	0.74	0.01	4.76	4.77	-0.02	0.00
Education & Training								
CD	19.17	22.84	-2.49	-1.18	19.73	23.53	-2.56	-1.24
DD	22.02	1.94	20.90	-0.82	22.74	2.54	20.90	-0.69
DDP	15.39	9.04	5.40	0.96	10.29	9.36	0.78	0.14
DDPS	8.00	6.78	1.09	0.13	7.32	7.35	-0.02	0.00
Personal Care								
CD	-0.79	-0.51	-0.28	0.00	-0.82	-0.53	-0.29	0.00
DD	2.20	-0.13	2.39	-0.06	2.20	-0.13	2.39	-0.06
DDP	0.34	-0.26	0.62	-0.02	-0.17	-0.26	0.09	0.00
DDPS	-0.15	-0.27	0.12	0.00	-0.29	-0.29	0.00	0.00
Purchasing Goods/Services								
CD	4.67	7.55	-2.39	-0.49	4.48	7.44	-2.45	-0.51
DD	22.36	2.34	20.06	-0.04	22.23	2.23	20.00	0.00
DDP	4.25	-1.02	5.18	0.10	-0.42	-1.18	0.75	0.01
DDPS	-1.49	-2.54	1.04	0.01	-2.58	-2.55	-0.02	0.00
Active Child Care								
CD	-9.09	-7.81	-1.40	0.11	-9.14	-7.82	-1.43	0.10
DD	14.21	0.45	11.72	2.04	14.30	0.53	11.66	2.12
DDP	0.77	-2.32	3.03	0.07	-2.23	-2.69	0.44	0.01
DDPS	-0.09	-0.69	0.61	0.00	-0.89	-0.87	-0.01	0.00
Housework								
CD	-11.49	-9.42	-2.26	0.18	-11.64	-9.51	-2.32	0.19
DD	20.77	1.43	18.93	0.41	20.63	1.31	18.95	0.37
DDP	4.53	-0.43	4.89	0.08	0.17	-0.55	0.71	0.01
DDPS	2.52	1.50	0.99	0.03	1.34	1.36	-0.02	0.00
Paid Work								
CD	6.74	2.95	3.69	0.11	6.86	2.96	3.79	0.11
DD	-31.43	-0.77	-30.90	0.24	-31.44	-0.78	-30.90	0.24
DDP	-7.74	0.25	-7.98	-0.02	-0.86	0.30	-1.16	0.00
DDPS	-1.87	-0.27	-1.61	0.00	-0.22	-0.26	0.03	0.00

4. SUMMARY AND RECOMMENDATIONS

Telephone time-use surveys have unique characteristics that make data collection more challenging. Unlike most other surveys, time-use surveys cannot accept proxy responses, so it is more likely that the probability of contacting a potential respondent is correlated with his or her activities. And because telephone time-use surveys ask respondents to report on their activities during the previous day, it is possible that the probability of interviewing the respondent about a given reference day will be correlated with the activities on that reference day. This paper shows how these characteristics can generate noncontact bias and activity bias. Two sets of computer simulations showed that the extent of these biases depends on the survey's strategy for contacting potential respondents.

In the first set of simulations, it was shown that the extent of the bias associated with any given contact schedule depends on the pattern of easy-to-contact (ETC) and hard-to-contact (HTC) days. The designated-day-with-postponement (DDP) schedule outperformed the other contact schedules for all of the activity patterns examined. These simulations also showed that estimates generated using a convenient-day (CD) schedule are sensitive to the within-person variance of the contact probability. Estimates of the time spent in activities that are positively correlated with the contact probability (for example, activities done at home) decrease as the variance increases. In contrast, estimates generated by other contact schedules are not sensitive to the within-person variance of the contact probability.

Given the results of the simple simulations, it is clear that the overall bias for the different contact strategies depends on the relative frequency of each pattern in the population. Direct data on these patterns do not exist, so the first set of simulations was augmented using CPS data on work schedules and actual time-use data from the 1992-94 EPA Time Diary Study. The results from the augmented simulations confirm those from the simple simulations, and show how the bias can affect estimates of time spent in specific activities. As expected, the CD contact strategy introduces systematic activity bias into time-use estimates. The time spent in activities done at home is underestimated, while time spent in activities done away from home is overestimated. There is no systematic activity bias in the samples generated by the DDP and DDPS strategies. The simulations also show that increasing the number of contact attempts reduces noncontact bias.

These results clearly show that the choice of contact strategy matters and point to two recommendations.

First, time-use surveys should use the DDP schedule. The DDP schedule generates less activity bias than the other contact schedules under all of the activity patterns tested. The DDPS schedule performed nearly as well in the more common activity patterns. But given that contact rates and field costs are a function of the number of contact attempts, the DDPS offers no cost advantage over the DDP

schedule. Hence, there is no reason to choose the DDPS schedule over the DDP schedule.

Second, time-use surveys need to take steps to minimize noncontact bias. Because noncontact bias is largely a function of the number of contact attempts, an obvious way to minimize noncontact bias would be to increase the number of contact attempts. No further elaboration will be made on this point, because other authors have looked at this issue in depth. For example, Bauman, Lavradas and Merkle (1993) show that age and employment status are related to the number of callbacks and that additional callbacks generate a more representative sample, and Botman, Massey and Kalsbeek (1989) propose a method for determining the optimal number of callbacks. Another alternative would be to try to increase the probability of contacting potential respondents. This could be done by determining when they are likely to be home and calling at those times, or by allowing them to call on their designated interview day. Paying incentives is another way to make potential respondents become "more available." A less costly approach to minimizing noncontact bias would be to adjust sample weights. Pothoff *et al.* (1993) show that, when the variable being measured is correlated (across individuals) with the contact probability, weighting based on the number of callbacks is practical and effective. In the end, the correct mix of these approaches will depend on the constraints facing the survey manager.

ACKNOWLEDGEMENTS

I thank John Eltinge, Mike Horrigan, Anne Polivka, Jim Spletzer, and Clyde Tucker for their comments and suggestions. The views expressed here are mine, and do not necessarily reflect those of the Bureau of Labor Statistics.

REFERENCES

- BAUMAN, S.L., LAVRADAS, P.J. and MERKLE, D.M. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 1070-1075.
- BOTMAN, S.L., MASSEY, J.D. and KALSBECK, W.D. (1989). Cost-efficiency and the number of allowable callbacks in the national health interview survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 434-439.
- HARVEY, A. (1993). Guidelines for time diary data collection. *Social Indicators Research*. 30, 197-228.
- HARVEY, A. (1999). Guidelines for time use data collection and analysis. *Time Use Research in the Social Sciences*, (Eds. W.E. Pentland, A.S. Harvey, P. Lawton and M.A. McColl). New York: Kluwer Academic/Plenum Publishers, 19-45.
- KALTON, G. (1985). Sample design issues in time diary studies. *Time, Goods, and Well-Being*, (Eds. F.T. Juster and F.P. Stafford). Ann Arbor: University of Michigan, Institute of Social Research, 333-351.

- KINSLEY, B., and O'DONNELL, T. (1983). Marking time: methodology report of the Canadian time use pilot study-1981. *Explorations in Time Use* (vol. 1), Ottawa: Department of Communications, Employment and Immigration.
- LAAKSONEN, S., and PÄÄKKÖNEN, H. (1992). Some methodological aspects on the use of time budget data. *Housework Time in Bulgaria and Finland*, (Eds. L. Kirjavainen, B. Anachkova, S. Laaksonen, I. Niemi, H. Pääkkönen and Z. Staikov). 86-104.
- LYBERG, I. (1989). Sampling, nonresponse, and measurement issues in the 1984-85 Swedish time budget survey. *Proceedings of the Fifth Annual Research Conference*: Department of Commerce, Bureau of the Census, 210-238.
- POTHOFF, R.F., MANTON, K.G. and WOODBURY, M.A. (1993). Correcting for nonavailability in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*. 88, 424, 1197-1207.
- STATISTICS CANADA (1999). *Overview of the Time Use of Canadians in 1998*, General Social Survey, Catalogue No. 12F0080XIE; Ottawa, Canada.
- STEWART, J. (2000). Alternative indexes for comparing activity profiles. Paper presented at the 2000 International Association for Time-Use Research Conference, Belo Horizonte, Brazil.
- TRIPLETT, T. (1995). Data Collection Methods for Estimating Exposure to Pollutants Through Human Activity Pattern Data: A National Micro-behavioral Approach. mimeo, Survey Research Center, University of Maryland.

Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples

ROBERT M. BELL and DANIEL F. MCCAFFREY¹

ABSTRACT

Linearization (or Taylor series) methods are widely used to estimate standard errors for the coefficients of linear regression models fit to multi-stage samples. When the number of primary sampling units (PSUs) is large, linearization can produce accurate standard errors under quite general conditions. However, when the number of PSUs is small or a coefficient depends primarily on data from a small number of PSUs, linearization estimators can have large negative bias. In this paper, we characterize features of the design matrix that produce large bias in linearization standard errors for linear regression coefficients. We then propose a new method, bias reduced linearization (BRL), based on residuals adjusted to better approximate the covariance of the true errors. When the errors are i.i.d., the BRL estimator is unbiased for the variance. Furthermore, a simulation study shows that BRL can greatly reduce the bias even if the errors are not i.i.d. We also propose using a Satterthwaite approximation to determine the degrees of freedom of the reference distribution for tests and confidence intervals about linear combinations of coefficients based on the BRL estimator. We demonstrate that the jackknife estimator also tends to be biased in situations where linearization is biased. However, the jackknife's bias tends to be positive. Our bias reduced linearization estimator can be viewed as a compromise between the traditional linearization and jackknife estimators.

KEY WORDS: Complex samples; Linearization; Jackknife; Satterthwaite approximation; Degrees of Freedom.

1. INTRODUCTION

Regression analysis of multi-stage samples has become very common in recent years (for example, Ellickson and McGuigan 2000; Shapiro, Morton, McCaffrey, Senterfitt, Fleishman, Perlman, Athey, Keeseey, Goldman, Berry and Bozzette 1999; Goldstein 1991; Landis, Lepkowski, Ekland and Stehouver 1982). Although hierarchical models (Bryk and Raudenbush 1992; Gelman, Carlin, Stern and Rubin 1995, Chapter 13) allow analysis of both fixed and random effects, many analysts prefer the simplicity of standard regression models when random effects are not of direct interest. Standard regression estimators produce unbiased parameter estimates that can be efficient, but the default standard error estimators do not account for the sample design, resulting in inconsistent standard errors (Kish 1965; Skinner 1989a). Various methods produce consistent standard error estimates applicable when the number of primary sampling units (PSUs) is sufficiently large. These include sample reuse methods such as the jackknife, bootstrap and balance repeated replication as well as linearization (or Taylor series) methods.

Linearization (Skinner 1989b) is a nonparametric method for estimating the standard errors of design-based statistics such as means and ratios as well as coefficients from linear and nonlinear regression models. By nonparametric, we mean that linearization does not rest on any assumptions about the within-PSU error structure, such as an assumption of constant intra-cluster correlation. When the number of PSUs can be considered large, linearization

produces consistent standard errors in the presence of multiple features of complex sample designs—stratification, multi-stage sampling, and sampling weights—as well as heteroskedastic errors (Fuller 1975). Because of these desirable properties and its increased availability in software such as SUDAAN, Stata, and SAS Version 8.0 (Shah, Barnwell, and Bieler 1997; StataCorp. 1999; SAS Institute, Inc. 1999), linearization has become a common method for estimating standard errors and confidence intervals and for conducting statistical tests on data from complex sample designs (for example, Ellickson and McGuigan 2000; Shapiro *et al.* 1999; Rust and Rao 1996). Linearization has also been proposed for estimating standard errors from Generalized Estimating Equations (GEE) fit to multi-stage data (Zeger and Liang 1986).

However, the linearization method has limitations. When the number of primary sampling units is small, standard error estimates can be severely biased low, they can have large coefficients of variation, and the standard degrees of freedom may be far too liberal (Kott 1994; Murray, Hannan, Wolfinger, Baker and Dwyer 1998). Consequently, standard linearization inference for coefficients based mainly on data from a small number of PSUs may produce confidence intervals that are too narrow and tests with Type I error rates that are substantially higher than their nominal values. Sample reuse methods like the jackknife have similar limitations.

In this paper, we characterize the design factors (*i.e.*, the distribution of explanatory variables within and between PSUs) that produce large bias in linearization and jackknife

¹ Robert M. Bell, Statistics Research Department, AT&T Labs-Research, Room C211, 180 Park Ave., Florham Park, NJ 07932; Daniel F. McCaffrey, Statistics Group, RAND, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516.

standard errors for linear regression coefficients and demonstrate that the problem can persist even when the number of PSUs is quite large. We then propose an alternative to the standard linearization estimator that is unbiased for independent, identically distributed (i.i.d.) errors and tends to greatly reduce bias otherwise. We also present approximate degrees of freedom for use with tests and confidence intervals based on our variance estimator. Simulation results show improved small sample properties of our alternative estimator and test compared with those of more traditional methods. Finally, we present an example of our methods using data from a national experiment evaluating care for depression.

2. BIAS OF THE LINEARIZATION METHOD

For simplicity, we restrict consideration in the body of this paper to unweighted linear regression for two-stage nonstratified samples. Extensions to weighted estimators and stratified samples are presented in McCaffrey, Bell and Botts (2001) and discussed further in section 8.

Let n equal the number of PSUs and m_i equal the number of final sampling units from the i -th PSU, for $i = 1, \dots, n$. The overall sample size is $M = \sum_i m_i$. We assume that $y_{ij} = \beta' x_{ij} + \varepsilon_{ij}$, where ε has mean 0 and covariance matrix \mathbf{V} , and where y_{ij} , x_{ij} , and ε_{ij} all refer to the j -th observation from the i -th PSU. We drop the standard OLS assumption of i.i.d. errors, assuming only that errors from distinct PSUs are uncorrelated. Specifically, we assume that \mathbf{V} is block diagonal, with $m_i \times m_i$ blocks \mathbf{V}_i for $i = 1, \dots, n$. In addition to the notation of this model, throughout the paper, we let \mathbf{I} denote an $M \times M$ identity matrix and \mathbf{I}_i equal an $m_i \times m_i$ identity matrix.

Let $\hat{\beta}$ denote the estimated coefficients of the linear regression model. To simplify presentation, we generally discuss a linear combination of the regression coefficients, $l' \hat{\beta}$, for an arbitrary column vector l . For the special case where one element of $l = 1$ and the rest are 0, $l' \hat{\beta}$ equals a single estimated coefficient. If errors are uncorrelated across PSUs, the variance of $l' \hat{\beta}$, is

$$\text{Var}(l' \hat{\beta}) = l' (\mathbf{X}' \mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}'_i \mathbf{V}_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l, \quad (1)$$

where \mathbf{X} and \mathbf{X}_i are the design matrices for the entire sample and for PSU i , respectively.

The standard linearization estimator of the variance of $l' \hat{\beta}$ is given by:

$$v_L = l' (\mathbf{X}' \mathbf{X})^{-1} \left(c \sum_{i=1}^n \mathbf{X}'_i \mathbf{r}_i \mathbf{r}'_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \quad (2)$$

where \mathbf{r}_i is the vector of residuals for the i -th PSU. Comparison of (1) and (2) shows that linearization simply involves estimating \mathbf{V}_i by a constant c times the outer product of the residuals. The constant c is typically set equal to $n/(n-1)$, the value used by SUDAAN and the Stata svy procedures (Shah, Barnwell, and Bieler 1997; StataCorp. 1999). For GEE procedures, Zeger and Liang (1986) set $c = 1$.

Under fairly general conditions, nv_L converges in probability to the variance of the asymptotic distribution of $\sqrt{n}(l' \hat{\beta} - l' \beta)$ and the relative bias of v_L is $O(1/n)$ as the number of PSUs gets large (Fuller 1975; Kott 1994). To demonstrate convergence for the bias of v_L , Kott (1994) assumes that the number of observations from every PSU is bounded and that elements of $(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ are bounded by B/n for a constant B . These assumptions effectively ensure that the influence of any PSU on the final estimate diminishes as the number of PSUs grows. Convergence of the bias of v_L holds for heteroskedastic data from stratified samples with unequal sampling weights and arbitrary correlation structure within PSUs. Unfortunately, consistency does not guarantee good properties for small to moderate numbers of PSUs.

Theorem 1. When $\mathbf{V} = \sigma^2 \mathbf{I}$ and $c = n/(n-1)$, $E(v_L) \leq \text{Var}(l' \hat{\beta})$ with equality if and only if $l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i$ is constant across i .

Proof. Without loss of generality, we assume that $\sigma^2 = 1$ so that $\mathbf{V} = \mathbf{I}$. The residual vector \mathbf{r} can be written as $(\mathbf{I} - \mathbf{H})\varepsilon$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ is the hat or projection matrix for \mathbf{X} . Thus, we have that $\mathbf{r}_i = (\mathbf{I} - \mathbf{H})_i \varepsilon$, where $(\mathbf{I} - \mathbf{H})_i$ contains the m_i rows of $(\mathbf{I} - \mathbf{H})$ for the i -th PSU. Consequently,

$$\begin{aligned} E(v_L) &= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\ &\quad \left(\sum_{i=1}^n \mathbf{X}'_i (\mathbf{I} - \mathbf{H})_i E(\varepsilon \varepsilon') (\mathbf{I} - \mathbf{H})_i' \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \\ &= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\ &\quad \sum_{i=1}^n \left(\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \quad (3) \end{aligned}$$

because $E(\varepsilon \varepsilon') = \mathbf{I}$ and $(\mathbf{I} - \mathbf{H})_i (\mathbf{I} - \mathbf{H})_i' = (\mathbf{I} - \mathbf{H}_{ii})$ for $\mathbf{H}_{ii} = \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i$. Let $\mathbf{D}_i = \mathbf{X}'_i \mathbf{X}_i - (1/n) (\mathbf{X}' \mathbf{X})$. Note that $\sum_i \mathbf{D}_i = \sum_i \mathbf{X}'_i \mathbf{X}_i - \mathbf{X}' \mathbf{X} = 0$. Thus,

$$\begin{aligned}
E(v_L) &= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
&\sum_{i=1}^n \left(\mathbf{X}'_i \mathbf{X}_i - [(1/n) \mathbf{X}' \mathbf{X} + \mathbf{D}_i] (\mathbf{X}' \mathbf{X})^{-1} [(1/n) \mathbf{X}' \mathbf{X} + \mathbf{D}_i] \right) \\
&\quad (\mathbf{X}' \mathbf{X})^{-1} l \\
&= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
&\quad \left(\mathbf{X}' \mathbf{X} - (1/n) \mathbf{X}' \mathbf{X} - \sum_{i=1}^n \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \\
&= l' (\mathbf{X}' \mathbf{X})^{-1} l - \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
&\quad \left(\sum_{i=1}^n \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \\
&= \text{Var}(l' \hat{\beta}) - \left(\frac{n}{n-1} \right) \left(\sum_{i=1}^n a'_i (\mathbf{X}' \mathbf{X})^{-1} a_i \right) \quad (4)
\end{aligned}$$

for $a_i = \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} l = [\mathbf{X}'_i \mathbf{X}_i - (1/n) (\mathbf{X}' \mathbf{X})] (\mathbf{X}' \mathbf{X})^{-1} l$. Because $(\mathbf{X}' \mathbf{X})^{-1}$ is positive definite, $E(v_L) \leq \text{Var}(l' \hat{\beta})$ with equality if and only if $a_i = 0$, or equivalently, $\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l$ is constant across the i .

Replication methods do not necessarily avoid the problem of bias for regression variance estimators. A jackknife estimator for multi-stage samples can be derived from the set of pseudo values $\{\hat{\beta}_{[i]}\}$, estimates of β from data that exclude the i -th PSU:

$$v_{JK} = [(n-1)/n] \sum_i l' (\tilde{\beta}_{[i]} - \hat{\beta}) (\tilde{\beta}_{[i]} - \hat{\beta})' l \quad (5)$$

(Cochran 1977; Rust and Rao 1996). If $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i , then

$$\begin{aligned}
v_{JK} &= [(n-1)/n] l' (\mathbf{X}' \mathbf{X})^{-1} \sum_i \mathbf{X}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \\
&\quad \mathbf{r}_i \mathbf{r}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l, \quad (6)
\end{aligned}$$

which follows from the updating formula $(\mathbf{X}' \mathbf{X} - \mathbf{X}'_i \mathbf{X}_i)^{-1} = (\mathbf{X}' \mathbf{X})^{-1} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1}$ (Cook and Weisberg 1982; Bell and McCaffrey 2002, page 34). Some authors (Efron and Tibshirani 1993) suggest an alternative jackknife estimator with $\hat{\beta}$ replaced by the mean of the $\tilde{\beta}_{[i]}$'s in (5). These two methods provide very similar estimates in our simulations, so we discuss only the version based on (5) in what follows.

Theorem 2. When $\mathbf{V} = \sigma^2 \mathbf{I}$ and $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i , then $E(v_{JK}) \geq \text{Var}(l' \hat{\beta})$ with equality if and only if $l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i$ is constant across i (proof in appendix).

The following example shows that the conditions for linearization and the jackknife estimators to be unbiased are

very restrictive even for simple linear regression.

Example 1. Consider simple linear regression. We have that

$$\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l = \frac{m_i}{Ms^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & s_i^2 + \bar{x}_i^2 \end{bmatrix} \begin{bmatrix} s^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} l$$

where s^2 and $\{s_i^2\}$ are ML estimates for the overall and within-PSU variances of x , with divisors M and $\{m_i\}$, respectively. So we have

$$\begin{aligned}
\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l &= \\
\frac{m_i}{Ms^2} \begin{bmatrix} s^2 + \bar{x}^2 - \bar{x}_i \bar{x} & \bar{x}_i - \bar{x} \\ (s^2 + \bar{x}^2) \bar{x}_i - (s_i^2 + \bar{x}_i^2) \bar{x} & s_i^2 + \bar{x}_i^2 - \bar{x}_i \bar{x} \end{bmatrix} l.
\end{aligned}$$

To have v_L and v_{JK} unbiased for the slope, *i.e.*, for $l' = (0, 1)$, we must have that $m_i(\bar{x}_i - \bar{x})$ and $m_i(s_i^2 + \bar{x}_i^2 - \bar{x}_i \bar{x})$ are both constant across i . The former implies that $\bar{x}_i \equiv \bar{x}$, and together they imply that $m_i s_i^2 = \sum_j (x_{ij} - \bar{x})^2$ is constant. Note that m_i need not be constant. These two conditions are not sufficient to guarantee unbiasedness for $l' = (0, 1)$, however. Additional algebra shows that the bias in the linearization estimator for the variance of the slope equals

$$-\frac{n}{(n-1)M^3 s^4} \left\{ \sum_{i=1}^n [m_i(\bar{x}_i - \bar{x})]^2 + \sum_{i=1}^n \left[\sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 - \bar{m} s^2 \right]^2 \right\}.$$

Consequently, the bias includes a part that is proportional to the weighted variance of the PSU means of x and another that is proportional to the variance of the within-PSU sums of squares.

The example shows that when the errors are i.i.d., v_L is unbiased only under very restrictive conditions. When $\mathbf{V} \neq \mathbf{I}$, Theorems 1 and 2 do not hold, and the bias in v_L can even be positive (see Example 2 of Bell and McCaffrey 2002).

In general, v_L tends to have negative bias. The estimator is the sum over PSUs of squares of linear combinations of residuals, $c^{1/2} l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{r}_i$. These sums of squares tend to be too small for two reasons: residuals are generally smaller than true errors due to overfitting, and residuals tend to have lower intra-cluster correlation than the errors. The factor $c = n/(n-1)$ corrects completely for these problems only in very restricted circumstances like the conditions in Theorem 1.

The bias of the linearization estimator (or the jackknife) increases with the between-PSU variance of the explanatory variables. Consequently, explanatory variables that are (nearly) constant within PSUs tend to exhibit the largest bias. When there are several such explanatory variables, there can be substantial underestimation of intra-cluster

correlations, leading to large bias in estimated variances for all the corresponding coefficients. Even greater bias potential appears to occur when certain PSUs account for most of the variability in the covariates and have disproportionate impact on the determination of $l' \hat{\beta}$.

3. THE BIAS REDUCED LINEARIZATION METHOD

Phillip Kott has proposed two methods for reducing the bias in linearization. Kott (1994) suggested correcting the bias in v_L by using the residuals and the design matrix to estimate the negative of the bias of v_L by \hat{R} ($\hat{R} > 0$, typically) and setting $v_{K94} = v_L / (1 - \hat{R}/v_L)$. Kott suggested the estimator v_{K94} rather than the more obvious $(v_L + \hat{R})$ as *ad hoc* compensation for the relative bias in \hat{R} as an estimator of the true negative bias, R .

In his 1996 paper, Kott suggests calculating the ratio of $\text{Var}(l' \hat{\beta})$ to $E(v_L)$ under the assumption that $\mathbf{V} = \mathbf{I}$ and adjusting v_L by the ratio. If $\mathbf{V} = \mathbf{I}$ then the resulting estimator v_{K96} will be unbiased.

In the context of generalized estimating equations, Mancl and DeRouen (2001) take a different approach to correcting the bias in the linearization estimator. They suggest adjusting the residuals from each PSU to reduce the bias in $\mathbf{r}_i \mathbf{r}_i'$ as an estimator of \mathbf{V}_i . For the unweighted linear model given in section 2, they approximate $E(\mathbf{r}_i \mathbf{r}_i')$ by $(\mathbf{I}_i - \mathbf{H}_{ii}) \mathbf{V}_i (\mathbf{I}_i - \mathbf{H}_{ii})$ and suggest replacing \mathbf{r}_i in $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{r}_i$ equation (2). Thus, for unweighted linear models the Mancl and DeRouen estimator equals $n/(n-1)v_{JK}$ and the properties on this estimator follow from the properties of the jackknife estimator.

We present an alternative approach that we first proposed in 1997 (McCaffrey and Bell 1997). The method is also based on replacing \mathbf{r}_i in equation (2) with adjusted residuals of the form $\mathbf{r}_i^* = \mathbf{A}_i \mathbf{r}_i$ intended to act more like the true errors ε_i . Like Kott (1996), we derive an estimator that eliminates the bias of v_L when \mathbf{V} equals \mathbf{U} , a specified block-diagonal covariance matrix, and reduces the bias for other \mathbf{V} . Like Mancl and DeRouen (2001) we adjust the residuals from each PSU. However, using \mathbf{U} we derive an alternative approximation to the $E(\mathbf{r}_i \mathbf{r}_i')$ and our resulting estimator is not proportional to the jackknife but rather can be seen as a compromise between the linearization and jackknife estimators. Our approach is also a generalization of the method of MacKinnon and White (1985), who adjust individual residuals to produce a heteroskedastically-consistent variance estimator (in the sense of White 1980) that is unbiased when the errors are independent and homoskedastic.

Theorem 3. For a specified block-diagonal covariance matrix \mathbf{U} , consider the class of estimators $v_L^* = l' (\mathbf{X}' \mathbf{X})^{-1} (\sum_{i=1}^n \mathbf{X}_i' \mathbf{A}_i \mathbf{r}_i \mathbf{r}_i' \mathbf{A}_i' \mathbf{X}_i) (\mathbf{X}' \mathbf{X})^{-1} l$, where \mathbf{A}_i satisfies $\mathbf{A}_i' (\mathbf{I}_i - \mathbf{H}_i) \mathbf{U} (\mathbf{I}_i - \mathbf{H}_i) \mathbf{A}_i' = \mathbf{U}_i$ for $i = 1, \dots, n$. If $\mathbf{V} = k\mathbf{U}$

for some scalar k , then $E(v_L^*) = \text{Var}(l' \hat{\beta})$.

Proof. The expected value of v_L^* is given by

$$E(v_L^*) = l' (\mathbf{X}' \mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}_i' \mathbf{A}_i (\mathbf{I}_i - \mathbf{H}_i) (k\mathbf{U}) ((\mathbf{I}_i - \mathbf{H}_i)' \mathbf{A}_i' \mathbf{X}_i) \right) (\mathbf{X}' \mathbf{X})^{-1} l = l' (\mathbf{X}' \mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}_i' (k\mathbf{U}_i) \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l = \text{Var}(l' \hat{\beta}).$$

Without external evidence to the contrary, an analyst is likely to use a working covariance matrix of the form $\mathbf{U} = \sigma^2 \mathbf{I}$, which simplifies the condition on \mathbf{A}_i to $\mathbf{A}_i' (\mathbf{I}_i - \mathbf{H}_{ii}) \mathbf{A}_i' = \mathbf{I}_i$ or

$$\mathbf{A}_i' \mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}. \quad (7)$$

We set $\mathbf{U} = \mathbf{I}$ in what follows.

A solution to equation (7) exists for PSU i whenever $(\mathbf{I}_i - \mathbf{H}_{ii})$ is full rank, which is true if all the eigenvalues of \mathbf{H}_{ii} are strictly less than 1 (the eigenvalues of \mathbf{H}_{ii} are always between 0 and 1). An eigenvalue of \mathbf{H}_{ii} may equal 1 – e.g., when the model includes a dichotomous explanatory variable that is one if and only if an observation falls in the i -th PSU.

For $m_i > 1$, \mathbf{A}_i is not unique. If \mathbf{A}_i satisfies $\mathbf{A}_i' \mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, then so does $\mathbf{O} \mathbf{A}_i$, for any $m_i \times m_i$ orthogonal matrix \mathbf{O} . If $\mathbf{V} = \sigma^2 \mathbf{I}$, the choice of \mathbf{A}_i is unimportant because any solution to (7) will produce an unbiased variance estimator. However, the resulting estimators are biased when $\mathbf{V} \neq \sigma^2 \mathbf{I}$, and the bias can vary greatly with the choice of \mathbf{A}_i . Heuristically, it makes sense to choose the solution \mathbf{A}_i “closest” to the identity matrix, so as to “mix” the residuals as little as possible. Two promising candidates are the Cholesky decomposition of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, which has all 0's below the diagonal, and the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$. Let \mathbf{P} be an orthogonal matrix whose columns are the eigenvectors of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ and $\mathbf{\Lambda}$ be a diagonal matrix containing the corresponding eigenvalues of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, so that $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}'$. Then for $\mathbf{\Lambda}^{1/2}$ equal to the elementwise square root of $\mathbf{\Lambda}$, $\mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}'$ is symmetric and solves (7). In contrast, multiplying either of these two solutions by a random orthogonal matrix could greatly distort the residuals.

Among the class of adjusted residuals of the form $\mathbf{A}_i \mathbf{r}_i$ where \mathbf{A}_i satisfies (7), those based on the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, $\mathbf{r}_i^* = \mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}' \mathbf{r}_i$, are “best” in the sense of Theil (1971) – i.e., they minimize the expected sum of the squared differences between the estimated and true i.i.d. errors (see pages 36-37 of Bell and McCaffrey 2002 for details). When there is intra-cluster correlation, simulation results in section 6 suggest that the bias of v_L^* based on the

symmetric square root is greatly reduced compared with that of the traditional linearization estimator, v_L . For these reasons, we consider only the symmetric root in the remainder of the paper and refer to the estimator using this root as the biased reduced linearization estimator, v_{BRL} .

As Kott (1994) proved for v_L , if the number of units in every PSU is bounded and the elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are bounded by B/n for some constant B (i.e., $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = O(1/n)$), then the bias in v_{BRL} is $O(n^{-2})$ and the relative bias is $O(1/n)$ (Bell and McCaffrey 2002, page 15).

4. VARIANCE OF THE ESTIMATORS AND TESTING

We note that v_L , v_{BRL} , and v_{JK} can all be written in the form

$$v^* = c l' (\mathbf{X}'\mathbf{X})^{-1} \sum_i \mathbf{X}_i' \mathbf{A}_i \mathbf{r}_i \mathbf{r}_i' \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l,$$

where: $c = n/(n-1)$, 1, or $(n-1)/n$, respectively, and $\mathbf{A}_i = \mathbf{I}_i$, $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}$, or $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, respectively. This formulation of the estimators shows that v_{BRL} can be viewed as a compromise between v_L and v_{JK} , chosen to offset their opposing biases.

Theorem 4. Let the error terms be distributed as multivariate normal with mean $\mathbf{0}$ and nonsingular covariance matrix \mathbf{V} . Then for any variance estimator of the form

$$v^* = c l' (\mathbf{X}'\mathbf{X})^{-1} \sum_i \mathbf{X}_i' \mathbf{A}_i \mathbf{r}_i \mathbf{r}_i' \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l,$$

v^* equals the weighted sum of independent $\chi^2_{\lambda_i}$ random variables where the weights are the eigenvalues of the $n \times n$ matrix for $\mathbf{G} = \{\mathbf{g}_i' \mathbf{V} \mathbf{g}_i\}$, for $\mathbf{g}_i = c^{1/2}(\mathbf{I} - \mathbf{H})_i' \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ (proof in appendix).

We can write v_L as a quadratic form $\mathbf{y}' \mathbf{G}^* \mathbf{y}$, where the M -by- M matrix $\mathbf{G}^* = \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i'$, so that v_L is a weighted sum of independent chi-square random variables with weights equal to the eigenvalues of $\mathbf{G}^* \mathbf{V}$. The proof consists of showing that the nonzero eigenvalues of $\mathbf{G}^* \mathbf{V}$ equal the nonzero eigenvalues of \mathbf{G} .

The mean and variance of v^* are simple functions of the eigenvalues of \mathbf{G} , namely $E(v^*) = \sum_{i=1}^n \lambda_i E(u_i^2) = \sum_{i=1}^n \lambda_i$ and $\text{Var}(v^*) = \sum_{i=1}^n \lambda_i^2 \text{Var}(u_i^2) = \sum_{i=1}^n 2\lambda_i^2$. If $\mathbf{V} = \sigma^2 \mathbf{I}$ and $\mathbf{X}_i' \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ for $i = 1, \dots, n$ are constant, conditions for v_L and v_{JK} to be unbiased, then Theorem 4 implies that av_L , av_{JK} , and av_{BRL} are all distributed χ^2_{n-1} for $a = (n-1)/\text{Var}(l' \beta)$ (Bell and McCaffrey 2002, pages 41-42). However, in general, the $\mathbf{X}_i' \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ will not be constant and the squared coefficient of variation will exceed $2/(n-1)$, the corresponding statistic for a χ^2_{n-1} random variable.

This excess variability is of particular concern when considering reference distributions for testing the null hypothesis that $l' \beta = 0$, with test statistics of the form $t = l' \hat{\beta} / \sqrt{v^*}$. For v_L , Shah, Holt and Folsom (1977)

suggested comparing t to a reference t -distribution with $n-1$ degrees of freedom, which is now the default in Stata (Stata Corp. 1999), SUDAAN (Shah, Barnwell and Bieler 1997) and SAS (SAS Institute 1999). The choice of $n-1$ degrees of freedom is motivated by the fact that v_L can be written as the sum of squares of n random variables $c^{1/2} l' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i' \mathbf{r}_i$. However, because the variance of $(n-1)v_L/E(v_L)$ tends to be greater than $2(n-1)$, tests that use a t -distribution with $n-1$ degrees of freedom would tend to have Type I error rates that exceed the nominal value, even if v_L were unbiased.

Satterthwaite (1946) suggested approximating the distribution of a linear combination of $\chi^2_{\lambda_i}$ variables by χ^2_f (up to a constant) where the first two moments of the linear combination match those of χ^2_f . We would approximate v_L , v_{BRL} or v_{JK} by a χ^2_f where $f = 2/cv^2 = (\sum_{i=1}^n \lambda_i)^2 / \sum_{i=1}^n \lambda_i^2$ and the λ_i are the eigenvalues of the corresponding matrix \mathbf{G} . Tests based on reference t -distributions with f degrees of freedom would be expected to provide better Type I error rates than tests based on $n-1$ degrees of freedom. Rust and Rao (1996) also suggest using a Satterthwaite approximation to estimate the degrees of freedom for the jackknife estimator. They present results for the estimator of a mean, while Theorem 4 extends this approach to testing linear combinations of regression coefficients. Kott (1994, 1996) suggests using the Satterthwaite approximation to estimate the degrees of freedom for tests based on his alternatives to linearization.

The coefficient of variation for any of the nonparametric variance estimators can be very large for certain designs. High variability occurs under the same conditions that v_L and v_{JK} are most biased – when residuals from only a few PSUs effectively determine the final variance estimate. This variability of the estimators is an inherent cost of using nonparametric techniques.

Because the Satterthwaite degrees of freedom f requires specifying the unknown matrix \mathbf{V} , we have investigated two methods for setting \mathbf{V} . The first treats \mathbf{V} as block-diagonal and estimates each block with the outer-product of the residuals for the PSU. Because preliminary simulation results showed that degrees of freedom based on this empirical estimate of \mathbf{V} produced tests that were extremely conservative, we do not present any simulation results for this method. Kott (1994) also found that estimating \mathbf{V} for use in the formula for estimated degrees of freedom proved unsatisfactory. Instead, we used a second method that sets \mathbf{V} identically equal to the identity matrix – i.e., it assumes independent, homoskedastic errors for purposes of determining degrees of freedom.

The distribution of v_{BRL} (and the other variance estimators) tends to be less skewed and have less mass in the lower tail than the distribution of a χ^2_f where f equals the Satterthwaite degrees of freedom. Hence, reference t -distributions based on the Satterthwaite approximation tend to overestimate tail probabilities. For example, when data from a couple of PSUs nearly determine the value of a

coefficient, the Satterthwaite degrees of freedom can be less than two, incorrectly implying a chi-square density that is infinite at zero. Consequently, the probability of very large t -statistics may not be as large as the Satterthwaite approximation would imply, especially when the Satterthwaite degrees of freedom are less than 4 or 5.

5. SIMULATION METHODS

We use a Monte Carlo simulation to study the properties of alternative variance estimators and tests for a balanced two-stage cluster sample with $n = 20$ PSUs and a constant $m = 10$ observations in each PSU. All simulation replications use a common design matrix \mathbf{X} with four explanatory variables chosen to represent a range of difficulty for nonparametric variance estimators. The first two explanatory variables, x_1 and x_2 , are dichotomous (0 or 1) and constant within PSU. The variable x_1 is 1 in half the clusters: 1, 3, ..., 19, while x_2 is 1 in just three clusters: 9, 10, and 11. Both x_3 and x_4 were generated from standard normal distributions. They differ in that x_3 was generated from a multivariate normal with intra-cluster correlation of 0.5 within PSU, while x_4 was generated from independent normal distributions. Observed intra-cluster correlations are 1.00, 1.00, 0.62 and -0.04, respectively. Observed correlations among the explanatory variables are all very small with the exception of $\text{Corr}(x_1, x_2) = 0.14$,

$\text{Corr}(x_1, x_3) = 0.25$ and $\text{Corr}(x_1, x_4) = -0.11$. The estimated regression coefficients are linear combinations of the dependent variable with multipliers given by the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which are shown in Figure 1. For the first three coefficients, and to a lesser extent $\hat{\beta}_3$, observations from the same PSU tend to have similar multipliers. Of more importance, $\hat{\beta}_2$, $\hat{\beta}_0$, and $\hat{\beta}_3$ are determined primarily by results in a small number of PSUs with relatively large multipliers (in absolute value). For example, Figure 1 shows that the multipliers for $\hat{\beta}_3$ are large for the second PSU, which has a mean that is over two standard deviations from the average PSU mean. In general, variance in the PSU means gives some PSUs greater weight for estimating $\hat{\beta}_3$.

The dependent variable was generated from the equation $y_{ij} = \beta'x_{ij} + \varepsilon_{ij}$, where $\beta = 0$ and the ε_i 's are standard multivariate normal random variables with intra-cluster correlation ρ . We use three alternative values of $\rho = 0, 1/9$, and $1/3$, corresponding to design effects for the sample mean of $\text{DEFF} = 1, 2$, and 4 , respectively ($\text{DEFF} = 1 + (m-1)\rho$). Monte Carlo results are based on 100,000 replications of \mathbf{y} for our fixed \mathbf{X} .

We evaluated the ordinary least squares (OLS) variance estimator, $s^2 l'(\mathbf{X}'\mathbf{X})^{-1}l$, and five nonparametric variance estimators: the standard linearization estimator given in equation (2) with $c = n/(n-1)$; the jackknife estimator given in (5); bias reduced linearization; and Kott's two adjustments to linearization. BRL and the Kott adjustments are all based on working intra-cluster correlations of $\rho = 0$.

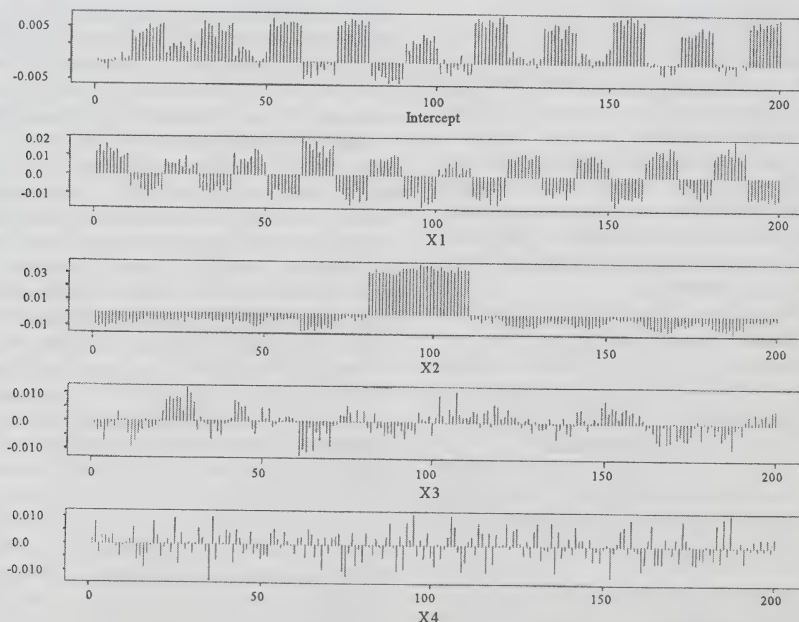


Figure 1. Values of the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for the design matrix used in simulations

We estimated Type I error rates for eight alternative test procedures based on 100,000 replications from the null hypothesis where each $\hat{\beta}_k = 0$, for $k = 0$ to 4. Each procedure compares a "t-statistic" against a reference t-distribution. For the t 's based on linearization, the jackknife, and BRL, we use critical values from t -distributions with both $(n - 1) = 19$ degrees of freedom and the corresponding Satterthwaite approximation. For Kott's methods, we use his proposed degrees of freedom. All computations were implemented in SAS.

6. SIMULATION RESULTS

Table 1 shows the bias of several variance estimators for the five regression coefficients (including the intercept) for $\rho = 0, 1/9$, and $1/3$. Except for Kott (1994), all values are exact based on the X matrix described above. Because Kott (1994) cannot be written as a linear functional, its bias is estimated from the Monte Carlo simulations, and the standard error of the bias is shown in parentheses.

Table 1

Bias of Variance Estimators (as a Percentage of the True Variance)

Estimator	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\rho = 0$					
OLS	0.0	0.0	0.0	0.0	0.0
Linearization	-9.6	-13.2	-32.5	-13.3	-1.8
Jackknife	11.7	17.2	51.2	17.6	2.1
Kott (1994)	4.0	2.5	-1.0	2.2	4.7
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	0.0	0.0	0.0	0.0	0.0
BRL	0.0	0.0	0.0	0.0	0.0
$\rho = 1/9$					
OLS	-50.2	-49.7	-50.7	-37.7	4.1
Linearization	-10.3	-14.2	-33.2	-17.1	-2.5
Jackknife	11.0	16.4	50.1	19.8	3.2
Kott (1994)	3.9	2.7	-0.8	1.5	4.6
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-0.8	-1.2	-1.0	-4.4	-0.7
BRL	-0.7	-1.0	-0.8	-1.2	0.1
$\rho = 1/3$					
OLS	-75.8	-75.5	-76.2	-65.3	13.8
Linearization	-10.7	-14.8	-33.5	-19.9	-4.1
Jackknife	10.7	15.9	49.5	21.4	5.9
Kott (1994)	3.6	2.4	-0.6	1.4	4.4
(Standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-1.2	-1.9	-1.5	-7.7	-2.3
BRL	-1.0	-1.5	-1.3	-2.1	0.4

Note: All values are exact except for Kott (1994), which is based on 100,000 simulation replications.

The OLS variances are unbiased for $\rho = 0$, but they are badly biased for $\rho = 1/9$ and $1/3$. As discussed in Wu, Holt and Holmes (1988), the OLS variances are too small by

roughly a factor of $1/[1 + \rho(m - 1)ICC_x]$, where ICC_x denotes the intra-cluster correlation for an x variable. Hence, for PSU-level variables (including the intercept), the OLS variances are too small by roughly a factor of $1/DEFF$. Similarly, the bias is smaller, but still substantial for x_3 , the individual-level variable with large intra-cluster correlation. The positive bias for the OLS variance of $\hat{\beta}_4$ results from the slight negative intra-cluster correlation for x_4 .

Linearization and the jackknife each suffer from large biases, relatively independent of ρ , but the biases point in opposite directions. For each estimator, the magnitude of the bias varies greatly among the coefficients. The largest biases (in absolute value) occur for $\hat{\beta}_2$, which depends mainly on the data from three PSUs. The next greatest biases occur for $\hat{\beta}_3$, followed closely by $\hat{\beta}_1$ and $\hat{\beta}_0$.

Except for $\hat{\beta}_4$, Kott (1994) has much smaller magnitude bias than linearization. However, the method tends to overcompensate, often resulting in notable positive bias. An exception is $\hat{\beta}_2$, for which Kott's estimator remains biased low.

By design, Kott (1996) and BRL eliminate the bias for $\rho = 0$. Consequently, choice among these alternatives should rest mainly on how well they hold down bias for $V \neq I$. Both methods reduce the magnitude of bias dramatically relative to linearization for $\rho = 1/9$ and $1/3$. Although differences between the two methods are often small, BRL does uniformly better, with its worst bias being -2.1 percent. While Kott (1996) is practically indistinguishable from BRL for the PSU-level variables, it performs substantially worse for $\hat{\beta}_3$ and $\hat{\beta}_4$.

The linearization, jackknife, BRL and Kott estimators are highly correlated with similar coefficients of variation. For any given regression coefficient, the correlation among the variance estimators always exceeded 0.969, with most exceeding 0.99 (not shown). The smallest correlations tended to be between the jackknife and other estimators. The coefficients of variation (also not shown) were largest for Kott (1994) and tended to be smallest for linearization and Kott (1996) (except for the intercept). For the intercept, the jackknife had the smallest coefficient of variation. The relative variance of the BRL estimator was similar to that of the alternative nonparametric methods. Its coefficient of variation was between 1 and 6 percent larger than that of the linearization estimator but about 5 to 10 percent smaller than that of Kott (1994). Thus, the five nonparametric variance estimators tend to differ from each other mainly by constant factors, and Table 1 summarizes the main difference among these variance estimators.

Table 2 shows the Satterthwaite degrees of freedom for each of the five coefficients for the linearization, jackknife, BRL and Kott variance estimators. For all estimators the degrees of freedom were calculated assuming $V = I$ and consequently depend only on the design matrix and not on the values of y . The approximations are similar for linearization and BRL although the linearization degrees of freedom tend to be slightly larger reflecting the fact that for

this design matrix the relative variances of the BRL estimators are marginally larger than those for linearization. Kott's approximation derives the coefficient of variation for a linearization-type estimator based on the true errors rather than the residuals. As a result, Kott's approximate degrees of freedom, which are larger than those for linearization or BRL, tend to overstate the precision of his estimator (see Kott 1994, section 6). Across all four estimators, the approximations are smallest for $\hat{\beta}_2$.

Table 2
Degrees-of-Freedom for Selected Estimators

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Satterthwaite (LIN)	9.02	14.45	3.30	11.56	16.65
Satterthwaite (Jackknife)	9.52	13.30	2.62	9.06	16.23
Satterthwaite (BRL)	9.24	14.08	2.90	10.26	16.45
Kott's method	10.33	16.41	4.32	11.36	17.44

Table 3 shows that Type I error rates for the standard linearization method with $(n-1)$ degrees of freedom consistently exceed 5 percent for all three values of p . Type I errors are most common for $\hat{\beta}_2$, where they reach as high as 16 percent, but they also occur much too frequently for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_3$, ranging from 7.0 to 8.8 percent. The magnitude of this problem correlates closely with the size of the bias of the linearization estimator (see Table 1). Type I error rates are much lower, 5.7 to 6.4 percent, for tests based on the Satterthwaite degrees of freedom. Thus using the alternative degrees of freedom improved the Type I error rates by about 30 to 88 percent.

There is a less consistent pattern for the Type I error probabilities for the jackknife. The jackknife with $(n-1)$ degrees of freedom tends to be conservative for $\hat{\beta}_1$ and $\hat{\beta}_3$, in accord with the positive bias in the jackknife variance. In contrast, the probability of Type I error is much too large for $\hat{\beta}_2$, and a bit too large in two of three cases for the intercept $\hat{\beta}_0$. The apparent explanation is that the choice of $(n-1)$ as the degrees of freedom for the reference t -distribution sometimes counteracts the bias in the jackknife variance. This conclusion is supported by the very low Type I error rates for the jackknife with Satterthwaite degrees of freedom; smaller degrees of freedom combined with large positive biases result in very conservative tests.

BRL with $(n-1)$ degrees of freedom improves substantially on linearization with the same degrees of freedom. Because BRL is unbiased when $p=0$, comparing the fifth row of the table against the first demonstrates the reduction in Type I errors that results from removing the bias of linearization. Excluding $\hat{\beta}_4$, BRL reduces Type I error rates by about 45 to 88 percent. However, BRL with $(n-1)$ degrees of freedom remains consistently liberal, especially for $\hat{\beta}_2$. Comparison of rows 2 and 5 of each section shows the relative impact of bias reduction and the Satterthwaite

adjustment. For $\hat{\beta}_0$ and $\hat{\beta}_2$, degrees of freedom are more important, while bias matters more for $\hat{\beta}_1$ and $\hat{\beta}_3$. Performance for BRL with the Satterthwaite approximation is very good, except for $\hat{\beta}_2$, where the Type I error falls to about 3 percent.

Table 3
Type I Error Rates for Tests of the Null Hypothesis that $\beta = 0$

Estimator	Df	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$p = 0$						
Linearization	$n-1$	7.54	7.00	15.99	7.35	5.38
Linearization	Satt	5.75	6.45	6.33	6.28	5.18
Jackknife	$n-1$	5.01	3.92	7.58	4.52	5.02
Jackknife	Satt	3.80	3.43	1.41	3.26	4.77
Kott (1994)	Kott	4.87	5.03	7.13	5.21	4.67
Kott (1996)	Kott	5.11	5.08	4.85	4.76	5.07
BRL	$n-1$	6.28	5.37	11.25	5.90	5.21
BRL	Satt	4.73	4.86	3.12	4.72	5.00
$p = 1/9$						
Linearization	$n-1$	7.81	7.14	16.19	8.18	5.34
Linearization	Satt	6.03	6.60	6.43	7.05	5.14
Jackknife	$n-1$	5.31	4.06	7.63	4.49	4.77
Jackknife	Satt	4.11	3.61	1.48	3.24	4.51
Kott (1994)	Kott	5.07	5.03	7.00	5.51	4.56
Kott (1996)	Kott	5.42	5.28	5.14	5.32	5.01
BRL	$n-1$	6.52	5.50	11.27	6.23	5.08
BRL	Satt	5.04	5.00	3.19	4.93	4.84
$p = 1/3$						
Linearization	$n-1$	8.10	7.28	16.39	8.79	5.66
Linearization	Satt	6.30	6.78	6.62	7.53	5.44
Jackknife	$n-1$	5.45	4.11	7.76	4.56	4.67
Jackknife	Satt	4.13	3.61	1.51	3.35	4.46
Kott (1994)	Kott	5.14	5.06	7.02	5.80	4.84
Kott (1996)	Kott	5.59	5.44	5.14	5.88	5.31
BRL	$n-1$	6.76	5.63	11.55	6.45	5.19
BRL	Satt	5.18	5.14	3.30	5.26	4.98

Note: Entries with a true value of 5.00 percent have standard errors of 0.07 percent.

Tests based on Kott's 1994 estimator with his proposed degrees of freedom perform very well for the coefficients where the variance estimator is biased upward. It appears that the upward bias in the variance estimator is offset by the upward bias in the approximate degrees of freedom. Kott's variance estimator is slightly negatively biased for $\hat{\beta}_2$ and therefore the upward bias in the degrees of freedom compounds the bias in the estimator resulting in a Type I error rate of about 7 percent for all three values of p .

Tests based on Kott's 1996 estimator also perform well. For almost all the coefficients and all values of p the Type I error rate is close to 5 percent. The exception is the test for $\hat{\beta}_3$ when $p = 1/3$, which has an error rate of 5.88 percent as a result of the moderate bias in the variance estimator.

7. EXAMPLE FROM THE PARTNERS IN CARE EXPERIMENT

We illustrate the methods in this paper using data from Partners in Care, a longitudinal experiment assessing the effect of "quality improvement" programs on care for depression in managed care organizations (MCOs) (Wells *et al.* 2000). The experiment followed 1356 patients who screened positive for depression in 1996-1997 in 43 clinics of seven MCOs. Clinics were assigned at random to one of three experimental cells: usual care, a quality improvement program supplemented by resources for medication follow-up, or a quality improvement program supplemented by resources for access to psychotherapists. Clinics were assigned at random after forming 27 clinic sets—three for each of nine blocks (six MCOs constituted single blocks, and one MCO was divided into three blocks based on ethnic mix of the clinics). Within blocks of more than three

clinics, clinic sets were combined to match as closely as possible on anticipated sample size and patient characteristics. See Wells *et al.* (2000) for additional details.

We present results from an OLS regression on the mental health summary score from the SF-12 (Ware, Kosinski and Keller 1995) for 1048 patients at 6-month follow-up. Scores were standardized to have mean 50 and standard deviation 10 in a general population, with higher scores indicating better health. As in Wells *et al.* (2000), the explanatory variable of primary interest is an intervention indicator that estimates the combined effect of medication or therapy versus care as usual. The first two columns of Table 4 show OLS coefficients and standard errors for the intervention effect and all the covariates used by, but not reported in, Wells *et al.* (2000). Our regression differs from theirs because we do not weight for nonresponse or impute for missing values of the outcome variable, but the results for the intervention effect agree reasonably closely.

Table 4
Comparison of OLS, Linearization, and BRL Inference for Partner-in-Care

Explanatory Variable	β_j	SE_{OLS}	SE_{LIN}	SE_{BRL}	DF_{BRL}	P-value		
			SE_{OLS}	SE_{OLS}		OLS	LIN	BRL
PSU-Level								
Intercept	28.795	3.409	1.03	1.06	23.7	0.000	0.000	0.000
Intervention	1.724	0.746	0.73	0.84	15.4	0.021	0.003	0.015
Block 1	1.386	1.867	0.63	0.80	2.7	0.458	0.244	0.426
Block 2	-0.031	1.576	0.88	1.07	3.6	0.984	0.982	0.986
Block 3	-1.042	1.230	0.53	0.61	3.9	0.397	0.117	0.241
Block 4	0.038	1.231	0.62	0.73	4.5	0.976	0.961	0.968
Block 5	-3.707	1.503	0.66	0.78	4.7	0.014	0.001	0.027
Block 6	-0.025	1.562	1.15	1.32	4.9	0.987	0.989	0.991
Block 7	-2.784	1.644	0.84	0.97	7.0	0.090	0.051	0.126
Block 8	0.822	1.233	0.93	1.03	12.0	0.505	0.476	0.527
Demographic								
Black	0.972	1.448	0.74	0.79	7.6	0.502	0.369	0.419
Hispanic	0.202	1.004	0.73	0.75	24.3	0.841	0.785	0.791
Other nonwhite	-1.033	1.409	0.77	0.80	21.6	0.463	0.349	0.369
Female	-0.502	0.803	1.09	1.12	23.1	0.532	0.571	0.581
Log of net worth + \$1,000	0.015	0.215	0.87	0.89	23.6	0.943	0.936	0.937
Less than high school	-1.690	1.217	1.00	1.04	25.3	0.165	0.173	0.192
Some college	-1.140	0.879	0.77	0.78	26.0	0.195	0.097	0.108
College graduate	-0.703	1.047	0.78	0.79	21.1	0.502	0.393	0.404
Age	0.059	0.032	0.91	0.93	26.5	0.064	0.047	0.056
Married	0.541	0.748	1.05	1.07	28.5	0.470	0.496	0.504
Baseline Health								
1 chronic condition (of 19)	-0.973	1.039	0.92	0.94	23.7	0.349	0.313	0.327
2 chronic conditions	0.198	1.116	0.87	0.90	23.0	0.859	0.840	0.846
3+ chronic conditions	-0.201	1.132	0.90	0.91	24.0	0.859	0.844	0.847
Depression and dysthymia	-5.305	1.335	0.93	0.95	25.8	0.000	0.000	0.000
Depression or dysthymia	-3.882	0.982	1.12	1.15	23.7	0.000	0.001	0.002
Prior depression only	-2.396	1.109	1.02	1.05	21.2	0.031	0.040	0.052
Mental component of SF-12	0.287	0.036	1.11	1.14	26.6	0.000	0.000	0.000
Physical comp of SF-12	0.079	0.036	0.88	0.89	24.6	0.029	0.017	0.022
Anxiety disorder	-2.438	0.749	1.20	1.23	26.3	0.001	0.010	0.014

Because patients from the same clinics could have similar outcomes, OLS standard errors could easily be too low—especially for PSU-level variables like Intervention. Columns 3 and 4 of Table 4 show the ratios of linearization and BRL standard errors to the OLS standard errors. We use clinic as the PSU because there is very little reason to expect correlations of errors across clinics after controlling for block.

Using the method of Wu, Holt and Holmes (1988), we estimate the intra-clinic correlation of the errors as -0.0026 , easily consistent with a true value of 0. Nonetheless, there is no reason to expect any of the correct standard errors to fall much below those obtained from OLS. Column 3 of Table 4 shows that the linearization standard errors frequently fall far below those obtained from OLS—especially for the PSU-level explanatory variables at the top of the table. Similarly, linearization with a reference t_{n-1} often produces much smaller P -values than does OLS. BRL improves over linearization. BRL standard errors are always larger and sometimes substantially larger than the linearization standard errors. For example, the BRL estimates for PSU-level explanatory variables are on average 15 percent larger than the linearization estimates. On the other hand, BRL standard errors for PSU-level variables are still often smaller than the OLS estimates. Thus, even though BRL estimators should be nearly unbiased, the variability in the estimators results in estimates for some coefficients that are small. The variability is also reflected in degrees of freedom that are very small for the block indicators and, while larger for patient level variables, are still considerably less than 42, the number of clusters minus one. The degrees of freedom are especially small, 7.6, for the indicator variable Black (equal to one if the patient was African American and zero otherwise). Plots analogous to Figure 1 show that Black was concentrated in three clusters. The Black indicator equals zero for all the patients in 24 of 43 clusters, and 48 of the 78 African Americans in the sample were found in just three clusters. As discussed in sections 2 and 4, the concentration of Black into a small number of clusters results in high variance for both estimators and large bias in the linearization estimator, both of which can be seen in Table 4.

8. DISCUSSION

Although linearization is a valuable tool that provides consistent standard errors and valid inference as the number of PSUs grows large in multi-stage samples, users should recognize problems with the method. Estimated variances of linear regression coefficients (including domain means) tend to be biased low—especially for coefficients (or linear combinations of coefficients) that depend largely on data from a small number of PSUs. Depending on the design, large biases can persist even when the total number of PSUs is quite large. The standard jackknife for multi-stage

samples tends to have at least as large bias in the opposite direction. Similarly, using a reference t distribution with degrees of freedom equal to one less than the number of PSUs may greatly understate the uncertainty in the estimated variance. Because the two problems (bias and overstated degrees of freedom) tend to occur in tandem for linearization, confidence intervals and statistical tests based on that method may be far too liberal.

Bias reduced linearization (BRL) produces unbiased variance estimates in the event that errors are homoskedastic and uncorrelated, and it tends to greatly reduce bias for other covariance structures investigated in our simulations. In our simulations, BRL consistently exhibited smaller biases than linearization by 90 percent or more and tended to improve substantially on Kott's 1994 adjusted linearization method. Results for BRL were comparable to those for Kott's 1996 method.

When BRL was used with the estimated Satterthwaite degrees of freedom, statistical inference improved greatly in comparison with the standard use of linearization. Bias reduction and Satterthwaite degrees of freedom seemed to contribute about equally to the improved performance. Although Satterthwaite's approximation may overcompensate, leading to conservative inference in certain situations, the problem does not seem noteworthy until the Satterthwaite degrees of freedom drop below 5 (based, in part, on simulations not reported in this paper). In such cases, analysts might choose to estimate critical values using simulations based on Theorem 4.

It is important to note some limitations of our simulation results. First, we only report results for four distinct explanatory variables plus an intercept. We choose those variables to span a wide variety of situations. Although some might describe x_2 as extreme or pathological, it is not outside the range of situations that we have seen in our own consulting work. Variables like x_2 can result from group-randomized trials (see section 7) or observational data where only a few PSUs exhibit a particular trait or from use of a series of dummy variables to represent levels of a categorical variable. Second, we present results only for $n = 20$ PSUs. To the extent that \mathbf{X} remains similar as n increases (e.g., by replication), Equation (4) implies that the bias declines in proportion to $1/(n-1)$. Also, the results observed for $n = 20$ could occur for much larger n if the bulk of the variation in \mathbf{X} is contributed by a few PSUs, and the determination of $l'\hat{\beta}$ depends similarly on a small number of PSUs. Finally, to reduce the number of factors affecting the results, we simplified the design in several ways: constant PSU sizes, no weights or strata, and little multicollinearity. We suspect that relaxing any of those constraints would actually tend to make standard linearization and the jackknife perform worse. We do not believe that the choice of $m = 10$ for the PSU size had much impact either way on our findings.

Although we believe that our proposed methods will prove valuable to analysts of multi-stage samples, these

methods will not completely solve the inference problem for unweighted linear regression. Both authors have frequently observed the disturbing situation where standard linearization methods produced shorter confidence intervals than methods that ignore the design. Certainly, the bias of v_L and improper use of $n-1$ degrees of freedom contribute to the frequency of this phenomenon, but our methods would not eliminate its occurrence (see section 7). Linearization, like sample reuse methods, necessarily produces estimators with high variance for some or possibly all coefficients in certain designs. When confronted with situations like the coefficients for our x_2 , where the Satterthwaite degrees of freedom fall near 3 or lower, analysts should seriously consider whether they can afford the large variability, and corresponding loss of power, that comes with nonparametric variance estimators. Parametric alternatives like hierarchical linear models or inference based on estimating a common intra-class correlation across all the PSUs (Wu, Holt, and Holmes 1988) should produce more stable results.

Although this paper has focused on unweighted linear regression for samples without stratification, we have no reason to expect that the bias and degrees-of-freedom problems of linearization would be lessened by stratification or for either weighted least squares or generalized linear models (GLMs). As shown in McCaffrey, Bell and Botts (2001) the BRL method extends immediately to weighted linear regression by using $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ in the main condition of Theorem 3. Because solutions to GLMs, such as logistic regression, are equivalent to the final steps of iteratively reweighted least squares (McCullagh and Nelder 1989), the obvious choice for these models is to use BRL based on the final weights and to set $\mathbf{U} = \mathbf{W}^{-1}$. Nevertheless, Theorem 3 does not extend to GLMs because the weights are estimated from the data, and we have not investigated the properties of BRL in this context.

Korn and Graubard (1995) suggest $v_L^{1/2}$ as a standard error estimator for stratified samples in situations where the stratification is non-informative. The same reasoning applies to $v_{\text{BRL}}^{1/2}$. Fuller (1975) proposed an alternative design consistent standard error estimator for stratified samples. Bell and McCaffrey (2002, pages 32-33) show that by adjusting the vector of residuals for each stratum, BRL can reduce or remove the model bias that can exist in Fuller's estimator.

ACKNOWLEDGEMENTS

We thank the referees and associate editor for valuable comments on an earlier draft. This work is supported in part by NSF Grant 0001763.

APPENDIX

Proofs of Theorems 2 and 4

Proof of Theorem 2

Following the first steps of the proof of Theorem 1, equation (6) implies that

$$E(v_{JK}) = \left(\frac{n-1}{n} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}_i' (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l.$$

The existence of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ implies that the eigenvalues of \mathbf{H}_{ii} are strictly less than 1, so that $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ can be written as $\sum_{j=0}^{\infty} \mathbf{H}_{ii}^j$. Consequently, letting $\mathbf{D} = (1/n)(\mathbf{X}'\mathbf{X})$ and $\mathbf{D}_i = (\mathbf{X}_i'\mathbf{X}_i) - \mathbf{D}$, we have

$$\begin{aligned} E(v_{JK}) &= \left(\frac{n-1}{n} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{k=1}^{\infty} [(\mathbf{D} + \mathbf{D}_i)(\mathbf{X}'\mathbf{X})^{-1}]^k \right) l \\ &= \left(\frac{n-1}{n} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{r=0}^{\infty} \sum_{s=0}^k \binom{k}{r} \frac{1}{n^{k-r}} [\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}]^r l \right) \\ &= \left(\frac{n-1}{n} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{r=0}^{\infty} \sum_{\substack{s=0 \\ r+s>0}}^{\infty} \binom{r+s}{r} \frac{1}{n^s} [\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}]^r l \right) \end{aligned}$$

The term for $r=0$ equals $l'(\mathbf{X}'\mathbf{X})^{-1} l = \text{Var}(l' \hat{\beta})$. The term for $r=1$ equals 0. By the binomial theorem,

$$\sum_{s=0}^{\infty} \binom{r+s}{r} \frac{1}{n^s} = \left(\frac{n}{n-1} \right)^{r+1},$$

so that the remaining terms can be paired, for $r=2, 4, 6, \dots$, to give

$$\begin{aligned} &\left(\frac{n}{n-1} \right)^r l' \sum_{i=1}^n \left\{ [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i]^{r/2} \right. \\ &\quad \left. [(\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1}] [\mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1}]^{r/2} \right\} l \end{aligned}$$

The middle factor in the summation can be written as ,

$$\left(\frac{n-2}{n-1} \right) (\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1} \right) (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}_i'\mathbf{X}_i) (\mathbf{X}'\mathbf{X})^{-1},$$

which is positive definite, so that the whole expression must be positive. Consequently, we have shown that $E(v_{JK}) \geq \text{Var}(l' \hat{\beta})$ with equality if and only if $l'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i = 0$, which is true if and only if $l'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i$ is constant across i .

Proof of Theorem 4

$$\begin{aligned} v^* &= c \sum_{i=1}^n l'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_i \mathbf{A}_i (\mathbf{I} - \mathbf{H})_i \varepsilon \varepsilon' (\mathbf{I} - \mathbf{H})'_i \\ &\quad \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l \\ &= \varepsilon' \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i' \varepsilon. \end{aligned}$$

Let \mathbf{P} equal the matrix of eigenvectors and \mathbf{A} denote the diagonal matrix with elements $\lambda_1, \dots, \lambda_M$ equal to the eigenvalues of $\mathbf{V}^{1/2} \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i' \mathbf{V}^{1/2} = \mathbf{B}'\mathbf{B}$ where $\mathbf{B}' = \mathbf{V}^{1/2} [\mathbf{g}_1 \mathbf{g}_2' \dots \mathbf{g}_n]$. Let $\mathbf{u} = \mathbf{P}' \mathbf{V}^{-1/2} \mathbf{y}$ where $\mathbf{V}^{1/2} \mathbf{V} \mathbf{V}^{-1/2} = \mathbf{I}$ defines $\mathbf{V}^{1/2}$, then the elements of \mathbf{u} are independent normal variables with variance 1 and

$$v^* = \mathbf{u}' \mathbf{A} \mathbf{u} = \sum_{i=1}^M \lambda_i u_i^2.$$

Let λ_i be any nonzero eigenvalue of $\mathbf{B}'\mathbf{B}$, then there exists a nonzero vector \mathbf{z} such that $\mathbf{B}'\mathbf{B}\mathbf{z} = \lambda_i \mathbf{z}$ and $\mathbf{B}\mathbf{B}'\mathbf{z} = \lambda_i \mathbf{z}$. Because $\mathbf{B}\mathbf{z} \neq \mathbf{0}$, λ_i is an eigenvalue of $\mathbf{B}\mathbf{B}'$. Similarly, any nonzero eigenvalue of $\mathbf{B}\mathbf{B}'$ is also an eigenvalue of $\mathbf{B}'\mathbf{B}$. Therefore, the nonzero eigenvalues of $\mathbf{B}'\mathbf{B}$ equal the nonzero eigenvalues of $\mathbf{B}\mathbf{B}' = \{\mathbf{g}_i' \mathbf{V} \mathbf{g}_i\}$.

REFERENCES

- BELL, R.M., and MCCAFFREY, D.F. (2002). *Bias Reduction in Linearization Standard Errors for Linear Regression with Multi-Stage Samples*. AT&T Labs-Research, Florham Park, NJ, TD-4S9H9T, www.research.att.com/~rbell.
- BRYK, A.S., and RAUDENBUSH, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition, New York: John Wiley & Sons Inc.
- COOK, R.D., and WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- EFRON, B., and TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- ELICKSON, P.L., and MCGUIGAN, K.A. (2000). Early predictors of adolescent violence. *American Journal of Public Health*. 90, 566-572.
- FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā C*, 37, 117-32.
- GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- GOLDSTEIN, H. (1991). Multilevel Modeling of Survey Data. *The Statistician*. 40, 235-244.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons Inc.
- KORN, E.L., and GRAUBARD, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A, General*. 158, 263-295.
- KOTT, P.S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology*. 20, 159-64.
- KOTT, P.S. (1996). Linear regression in the face of specification error: model-based exploration of randomization-based techniques. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. 39-47.
- LANDIS, J.R., LEPKOWSKI, J.M., EKLAND, S.A. and STEHOUWER, S.A. (1982). A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and Health Statistics, Series 2*, 92, Washington, D.C: US Government Printing Office.
- MACKINNON, J.G., and WHITE, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*. 29, 305-325.
- MANCL, L.A., and DEROUEN, T.A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*. 57, 126-134.
- MCCAFFREY, D.F., and BELL, R.M. (1997). Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units. Paper presented at the Joint Statistical Meetings, Anaheim CA.
- MCCAFFREY, D.F., BELL, R.M. and BOTTS, C.H. (2001). Generalizations of bias reduced linearization. *Proceeding of the Survey Research Methods Section*, American Statistical Association.
- MCCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*. Second Edition, London: Chapman and Hall.
- MURRAY, D. M., HANNAN, P. J., WOLFINGER, R. D., BAKER, W.L. and DWYER, J.H. (1998). Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine*. 17, 1581-1600.
- RUST, K.F., and RAO, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*. 5, 283-310.
- SAS INSTITUTE INC. (1999). *SAS/STAT* User's Guide, Version 8*. Cary, NC: Author.
- SATTERTHWAIT, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*. 2, 110-114.
- SEARLE, S.R., CASELLA, G. and MCCULLOCH, C.E. (1992). *Variance Components*. New York: John Wiley & Sons Inc.
- SHAH, B.V., BARNWELL, B.G. and BIELER, G.S. (1997). *SUDAAN User* Manual, Release 7.5*. Research Triangle Park, NC: Research Triangle Institute.
- SHAH, B.V., HOLT, M. M. and FOLSOM, R.E. (1977). Inference About Regression Models from Survey Data. *Bulletin of the*

International Statistical Institute. 41, 43-57.

- SHAPIRO, M.F., MORTON, S.C., MCCAFFREY, D.F., SENTERFITT, J.W., FLEISHMAN, J.A., PERLMAN, J.F., ATHEY, L.A., KEESEY, J.W., GOLDMAN, D.P., BERRY, S.H. and BOZZETTE, S.A. (1999). Variations in the care of hiv-infected adults in the United States; results from the hiv cost and services utilization study. *Journal of the American Medical Association.* 281, 2305-2315.
- SKINNER, C.J. (1989a). Introduction to Part A. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, and T.M.F. Smith). New York: John Wiley & Sons Inc. 23-57.
- SKINNER, C.J. (1989b). Domain means, regression and multivariate analyses. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons Inc. 59-88.
- STATACORP. (1999). *Stata Statistical Software: Release 6.0*. College Station, TX: Author.
- THEIL, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons Inc.
- WARE, J.E., JR., KOSINSKI, M. and KELLER, S.D. (1995). *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales*. Boston, Mass: The Health Institute, New England Medical Center.
- WELLS, K.B., SHERBOURNE, C., SCHOENBAUM, M., DUAN, N., MEREDITH, L., UNUTZER, J., MIRANDA, J., CARNEY, M. and RUBENSTEIN, L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association.* 283, 212-220.
- WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica.* 48, 817-838.
- WU, C.J.F., HOLT, D. and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association.* 83, 150-9.
- ZEGER, S.L., and LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 42, 121-130.

Design Effects of Sampling Frames in Establishments Survey

MONROE G. SIRKEN¹

ABSTRACT

When stand-alone sampling frames that list all establishments and their measures of size are available, establishment surveys typically use the Hansen-Hurwitz (HH) pps estimator to estimate the volume of transactions that establishments have with populations. This paper proposes the network sampling (NS) version of the HH estimator as a potential competitor of the HH estimator. The NS estimator depends on the population survey-generated establishment frame that lists households and their selection probabilities in a population sample survey, and the number of transactions, if any, of each household with each establishment. A statistical model is developed in this paper to compare the efficiencies of the HH and NS estimators in single-stage and two-stage establishment sample surveys assuming the stand-alone sampling frame and the population survey-generated frame are flawless in coverage and size measures.

KEY WORDS: Stand-alone establishment frames; Population survey-generated establishment frames; Hansen-Hurwitz estimator; Network sampling estimator.

1. INTRODUCTION

Listings of establishments that have transactions with households in population sample surveys serve as sampling frames of establishment surveys whenever the transactions reported by households in the population surveys are matched with the records of their establishments. For example, the listings of establishments that have transactions with households in the National Medical Expenditure Panel Survey (MEPS), a national population sample survey, serve as sampling frames for medical provider surveys that supplement and verify the medical expenditures of the transactions reported by MEPS household respondents (Cohen 1998). However, listings of establishments that have transactions with households in population sample surveys rarely serve as frames of establishment surveys that collect information about the transactions that establishments have with all households. The Current Price Index (CPI) produced by the Bureau of Labor Statistics is a notable and rare exception of a Federal establishment survey that depends on a population survey-generated sampling frame. The CPI Pricing Survey, a national retail establishment survey, that collects prices for a basket of consumer goods purchased by all customers, uses as its sampling frame the listings of retail establishments that have transactions with households in the CPI Continuing Point of Purchase Survey. (Leaver and Valliant 1995).

After reviewing plans of the National Center for Health Statistics (NCHS) to restructure its family of independent national surveys of health providers (hospitals, physicians, clinics, *etc.*), a Panel of the Committee on National Statistics proposed (Wunderlich 1992) using listings of health care providers reported by households in the National Health Interview Survey (NHIS), an ongoing

national household sample survey (Massey, Moore, Parsons and Tadros 1991) as the sampling frames for national surveys of health care providers. The Committee thought that, especially in the current environment of rapid changes in listings of health care providers due to rapid changes in the nation's health care delivery system, the NHIS-generated health care provider frames would be more accurate and easier and less expensive to construct and maintain than the free-standing health care provider frames currently in use. Soon after the Panel report was issued, NCHS initiated a research project on population survey-generated sampling frames that is briefly summarized below.

Initially, the research focused almost exclusively on the statistical properties of NHIS-generated frames of health care providers. Judkins, Berk, Edwards, Mohr, Stewart and Waksberg (1995) studied the quality of the free-standing health provider frames currently in use or of potential use, and discussed the kinds of medical providers for which NHIS-generated frames would seem to have the greatest potential. Subsequently, Judkins, Marker, Waksberg, Botman and Massey (1999) made rough comparisons of the efficiencies of dental surveys using the NHIS-generated sampling frame and using the free-standing frame, and concluded that NHIS-generated health care provider frames deserve serious consideration whenever reasonably complete free-standing health care provider frames with reasonably good size measures are unavailable.

In recent years, the research has focused on the statistical properties of estimators that depend on population-generated sampling frames and has become more theoretically focused than formerly. The conceptual difficulties initially encountered in developing unbiased estimators for the population survey-generated frame because the same establishments have transactions with multiple households were overcome by applying network sampling theory. (Sirken

¹ Monroe G. Sirken, Senior Research Scientist, National Center for Health Statistics, U.S.A.

1997; Thompson 1992). Sirken, Shimizu and Judkins (1995) developed the network sampling version of the HH estimator, referred to in this paper as the NS estimator, and Sirken and Shimizu (1999) developed the network sampling version of the Horwitz-Thompson (HT) estimator. This paper develops a statistical error model that compares the efficiencies of the NS estimator that depends on the population survey-generated frame, and the HH estimator that depends on the free-standing frame. The error model assumes both frames are flawless in establishment coverage and size measures and have equivalent construction and maintenance costs. Though the model assumes a srs design for the population survey that generates population survey-generated sampling frame, the model can be applied to other kinds of population survey designs that are not considered in this paper.

This paper is organized as follows. Notation follows in section 2. Section 3.1 and section 3.2 respectively present the pps self-weighted HH estimator and variance of the two-stage establishment sample survey that depends on the free-standing sampling frame, and the NS estimator and variance of a two-stage establishment survey that depends on the population survey-generated frame. The error model is developed in sections 4.1-4.4. The difference between two-stage HH and NS variances of equivalent expected sample sizes is developed in section 4.1. In section 4.2, the first stage variance component of the two-stage NS estimator is split into variance components representing effects of households with and without transactions, and section 4.3 shows the design effects of the NS estimator in single stage sampling. Second stage variance components of the NS and HH estimators are compared in section 4.4. In the concluding section 5, the error model's major findings comparing efficiencies of HH and NS estimators in single-stage and two-stage establishment surveys are briefly summarized, and limitations of the model are briefly discussed. The appendix presents the proof of a statistical statement appearing in section 4.2.

2. NOTATION

Let N_j = the number of households having transactions with establishment j ($j = 1, 2, \dots, R$), N_o = the number of households not having transactions with any establishments, and N^* = the number of distinct households having transactions with R establishments. Then, $N = N^* + N_o$ = the total number of households.

Let M_{ij} = the number of transactions of establishment j ($j = 1, 2, \dots, R$) with household i ($i = 1, 2, \dots, N$), where $M_{ij} \geq 0$ when establishment j has transactions with household i , and $M_{ij} = 0$ when establishment j and household i do not have transactions. Then, $M_j = \sum_{i=1}^N M_{ij}$ = the number of transactions of establishment j with N households, and $M = \sum_{j=1}^R M_j$ = the number of transactions of M establishments with N households, and $\bar{M} = M/N$ the average number of transactions per household.

Let X_{jk} denote the value of the x -variate for transaction k ($k = 1, \dots, M_j$) of establishment j ($j = 1, 2, \dots, R$). Then, $X_j = \sum_{k=1}^{M_j} X_{jk}$ = the sum of the x -variate over the M_j transactions of establishment j , and $X = \sum_{j=1}^R X_j$ = sum of the x -variate over the M transactions of R establishments. Let $\bar{X}_j = X_j/M_j$ = the average value of the x -variate over the M_j transactions of establishment j , and $\bar{X} = X/M$ = the average value of the x -variate over M transactions.

3. ESTIMATORS AND VARIANCES

3.1 The HH Estimator and Variance

Consider a two-stage self weighted establishment sample survey using a free-standing establishment sampling frame that lists all R establishments and their measures of size, M_j ($j = 1, 2, \dots, R$). Establishments are the primary sampling units (psu's), and transactions are the secondary sampling units. A sample of r establishments is selected with pps with replacement from the free-standing frame, and a sample of size $t_{HH} < \min(M_1, \dots, M_j, \dots, M_R)$ transactions each, where t_{HH} is a positive integer, is independently selected by simple random sampling without replacement for each sample establishment j ($j = 1, 2, \dots, r$).

The unbiased self-weighted pps HH estimator of X is

$$X'_{HH} = \frac{M}{r} \sum_{j=1}^r \bar{X}'_j \quad (1)$$

where $\bar{X}'_j = \sum_{k=1}^{t_{HH}} X_{jk}/t_{HH}$ is the unbiased estimate of $\bar{X}_j = X_j/M_j$ ($j = 1, 2, \dots, R$). Because establishments are selected with replacement, the HH estimator counts \bar{X}_j as many times as establishment j is selected in the sample.

The variance of the X'_{HH} is (Thompson 1992)

$$\text{Var}(X'_{HH}) = \frac{M^2}{r} \sigma_{HH1}^2 + \frac{M}{rt_{HH}} \sum_{j=1}^R (M_j - t_{HH}) \sigma_j^2 \quad (2)$$

where the first and second terms respectively on the right side of (2) are the first and second stage variance components, and

$$\sigma_{HH1}^2 = \frac{1}{M} \sum_{j=1}^R M_j (\bar{X}_j - X/M)^2 \quad (3)$$

is the between establishment population variance, and

$$\sigma_j^2 = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (X_{jk} - X_j/M_j)^2 \quad (4)$$

is the within establishment population variance of establishment j .

3.2 The NS Estimator and Variance

Consider a two-stage establishment sample survey that depends on a population survey-generated frame. The frame lists n sample households H_i ($i = 1, 2, \dots, n$) that were

enumerated in a population sample survey. For each listed household H_i' , the frame provides π_i , its selection probability in the household survey, and M_{ij} , the number of its transactions with each distinct establishment j ($j = 1, 2, \dots, R$). (The M_{ij} 's are reported by household respondents in the population sample survey).

Each of the n listed households in the population survey-generated frame represents a cluster of establishments ranging in size from 0 to R establishments with whom the household has transactions. The n clusters of establishments are the primary sampling units, and the M_j ($j = 1, 2, \dots, r$) transactions of the r sampled establishments are secondary sampling units. The transaction sample for establishment j ($j = 1, 2, \dots, R$) is selected as follows: a srs sample of size $t_{NS} M_{ij} < \min(M_1, M_2, \dots, M_r)$ transactions is independently selected without replacement for each sample household H_i' ($i = 1, 2, \dots, n$), where t_{NS} is a positive integer. The transaction sample size of establishment j ($j = 1, 2, \dots, R$) is equal to $t_{NS} \sum_{i=1}^n M_{ij}$, and the total transaction sample size is equal to τt_{NS} , where $\tau = \sum_{i=1}^n \sum_{j \in A_i} M_{ij}$ = the sum of the transactions over n sample households is a random variable.

The NS estimator of X is

$$X'_{NS} = \sum_{i=1}^n \frac{1}{\pi_i} \sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$$

where A_i is the cluster of distinct establishments that have transactions with sample household H_i' , and

$$\bar{X}'_j(i) = \sum_{k=1}^{t_{NS} M_{ij}} X_{jk} / (t_{NS} M_{ij})$$

is an unbiased estimate \bar{X}_j for a sample of $t_{NS} M_{ij}$ transactions of establishment j . Because households are selected with replacement, the NS estimator counts the quantity $\sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$ every time household H_i ($i = 1, 2, \dots, n$) is selected in the sample, and because the same establishment has transactions with multiple households, the NS estimator counts the quantity $M_{ij} \bar{X}'_j(i)$ every time a sample household i ($i = 1, 2, \dots, n$) contains establishment j .

Assuming a srs design in the population survey, $\pi_i = n/N$, and the network sampling estimator is

$$X'_{NS} = \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j(i). \quad (5)$$

The NS estimator is an unbiased estimator of X .

$$\begin{aligned} E(X'_{NS}) &= \sum_{i=1}^n E \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j \\ &= \sum_{i=1}^R M_j \bar{X}_j = \sum_{j=1}^R X_j = X. \end{aligned}$$

The NS estimator in (5) is self-weighted because we have assumed that the n households are selected by srs. It would be a self-weighted estimator whenever the sample design of the population sample survey that generates the establishment sampling frame is self-weighted. When $N = N^* = M$, implying that N^* households each has a single transaction, and $N_0 = N - N^*$ households are without transactions, and when $n = r$ and $t_{NS} = t_{HH}$, the HH and NS estimators are equivalent.

$$\begin{aligned} X'_{NS} &= \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} \bar{X}'_j \\ &= \frac{M}{r} \sum_{j=1}^R \bar{X}_j = X_{HH}. \end{aligned} \quad (6)$$

The variance of the NS estimator (5), under srs sampling with replacement of n households and independent selections of $t_{NS} M_{ij}$ transaction by srs without replacement for each establishment j linked to household H_i , is (Sirken *et al.* 1995)

$$\begin{aligned} \text{Var}(X'_{NS}) &= \frac{N^2}{n} \sigma_{NS1}^2 + \frac{N}{n t_{NS}} \sum_{i=1}^n \sum_{j=1}^R \\ &\quad M_{ij} \frac{M_j - t_{NS} M_{ij}}{M_j} \sigma_j^2 \end{aligned} \quad (7)$$

where the first and second terms respectively on the right side of (7) are the first and second stage variance components, and

$$\sigma_{NS1}^2 = \frac{1}{N} \sum_{i=1}^n \left(\sum_{j \in A_i} M_{ij} \bar{X}'_j - X/N \right)^2 \quad (8)$$

is the population variance between households, and σ_j^2 , the population variance within establishment j as defined in (4). An unbiased estimate of NS variance is

$$\text{Var}(X'_{NS}) = \frac{N^2}{n(n-1)} \sum_{i=1}^n \left[\sum_{j \in A_i} M_{ij} \bar{X}'_j(i) - \bar{X}' \right]^2 \quad (9)$$

where $\bar{X}' = X'/N$.

4. THE ERROR MODEL

4.1 HH and NS Variances of Equivalent Expected Sample Size

Subtracting (2) from (7), the difference between the variances of the HH and NS estimators of X is

$$\begin{aligned} \text{Var}(X'_{\text{NS}}) - \text{Var}(X'_{\text{HH}}) = & \left[\frac{N^2}{n} \sigma_{\text{NS1}}^2 - \frac{M^2}{r} \sigma_{\text{HH1}}^2 \right] \\ & + \left[\frac{N}{nt_{\text{NS}}} \sum_{i=1}^N \sum_{j=1}^R M_{ij} \frac{M_j - t_{\text{NS}}}{M_j} M_{ij} \sigma_j^2 \right. \\ & \left. - \frac{M}{rt_{\text{HH}}} \sum_{j=1}^R (M_j - t_{\text{HH}}) \sigma_j^2 \right] \quad (10) \end{aligned}$$

where the first and second set of bracketed terms respectively on the right side of (10) represent the differences between the primary and secondary variance components of the HH and NS estimators of X .

Let $m_{\text{NS}} = \tau t_{\text{NS}}$ = the size of the transaction sample in the establishment survey using the population survey-generated frame, where t_{NS} , a positive integer, is the size of the transaction sample selected per transaction of the n sample households, and $\tau = \sum_{i=1}^n \sum_{j \in A_i} M_{ij}$ = sum of the transactions of n sample households.

Clearly, τ is a random variable and its expected value conditional over all samples of n households is $E(\tau|n) = n\bar{M}$ where $\bar{M} = M/N$ = average household transaction size. It follows that $E(m_{\text{NS}}|n) = t_{\text{NS}} E(\tau|n) = n\bar{M} t_{\text{NS}}$ is the expected transaction sample size of the NS estimator conditional over all samples of n households.

Let $m_{\text{HH}} = r t_{\text{HH}}$ = the size of the transaction sample in the establishment survey using the stand-alone frame, where r = the establishment sample size, and t_{HH} = the transaction sample size per selected establishment. Let $r = E(\tau|n) = n\bar{M}$ and let $t_{\text{HH}} = t_{\text{NS}} = t$, and it follows the expected transaction sample sizes of the NS and HH estimators conditional over all samples of n households are equivalent, namely, $E(m_{\text{HH}}|n) = t E(\tau|n) = nt\bar{M} = E(m_{\text{NS}}|n)$.

Calibrating the establishment and transaction sample sizes in this manner assures that HH and the NS establishment surveys are conducted under roughly the same fiscal constraints if per establishment and per transaction field costs are about the same in both surveys. It is noteworthy, however, that this cost equation does not take into account the differences in costs between constructing and maintaining stand-alone establishment frames and population survey-generated establishment frames.

Substituting $r = n\bar{M}$, $t_{\text{HH}} = t_{\text{NS}} = t$, and $M = N\bar{M}$ in formula (9), the difference between the NS and HH variances of equivalent expected establishment and transaction sample size conditional over all samples of n households is

$$\begin{aligned} \text{Var}(X'_{\text{NS}}) - \text{Var}(X'_{\text{HH}}) = & \frac{N^2}{n} [\sigma_{\text{NS1}}^2 - \bar{M} \sigma_{\text{HH1}}^2] \\ & - \frac{N}{nt} \sum_{j=1}^R \sigma_j^2 [(M_j - t) - \sum_{i=1}^N \frac{M_{ij}(M_j - M_{ij})}{M_j}] \quad (11) \end{aligned}$$

The first term and second terms respectively on the right side of (11) represent the difference between the first stage and second stage variance components of the NS and HH estimators of equivalent expected sample sizes conditional over all samples of n households.

4.2 Decomposition of the Single Stage NS Population Variance

Typically, some households do not have transactions with any establishments, and the percentage varies by type of establishment. For example, medical care utilization by families in the United States varies greatly by type of health care provider (Dicker and Sunshine 1987). During a 12 month period, 70 percent of families were not admitted to hospitals, 7 percent did not have ambulatory physician visits, and 28 percent did not have dental visits.

Let

$$P = \frac{N^*}{N} = \text{fraction of } N \text{ households with one or more transactions, and}$$

$$P_0 = 1 - P \frac{N_0}{N} = \text{fraction of } N \text{ households without any transactions.}$$

We demonstrate in the Appendix that the single stage population variance of the NS estimator of X , when expressed as a function of P , decomposes into 2 parts

$$\begin{aligned} \sigma_{\text{NS1}}^2(P) &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N} \right)^2 \\ &= P \sigma_{\text{NS1}^*}^2 + \sigma^2(P) E_{\text{NS1}^*}^2, \quad 0 < P \leq 1 \quad (12) \end{aligned}$$

where

$$\sigma_{\text{NS1}^*}^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 \quad (13)$$

is the single stage population variance of the x -variate over the truncated population of N^* households with one or more transactions,

$$E_{NS1}^2 = \left(\frac{X}{N^*} \right)^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j \right)^2 - \sigma_{NS1}^2 \quad (14)$$

is the expected value squared of the x -variate over the truncated population of N^* households and

$$\sigma^2(P) = P(1-P) \quad (15)$$

is the variance of the binomial variable P . For fixed M , the function $\sigma_{NS1}^2(P|M)$ is maximum when

$$P = P_{\max} = \frac{1}{2} \left[(\sigma_{NS1}^2 / E_{NS1}^2) + 1 \right] \leq 1.$$

If $\sigma_{NS1}^2 \geq E_{NS1}^2$, $P_{\max} = 1$ and if $\sigma_{NS1}^2 < E_{NS1}^2$, $1/2 < P_{\max} < 1$.

When $P = 1$, $\sigma^2(P = 1) = 0$ and therefore $\sigma_{NS1}^2(P = 1) = \sigma_{NS1}^2$. If $P = M = (M/N) = 1$, implying that each of N households has a single transaction,

$$\sigma_{NS1}^2(P = \bar{M} = 1) = \sigma_{NS1}^2(N^* = M) = \sigma_{HH1}^2 \quad (16)$$

because

$$\begin{aligned} \sigma_{NS1}^2(N^* = M) &= \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 \\ &= \frac{1}{M} \sum_{j=1}^R M_j \left(\bar{X}_j - \frac{X}{M} \right)^2 = \sigma_{HH1}^2 \quad (17) \end{aligned}$$

and, $\sigma^2(P = 1) = 0$. In other words when $P = \bar{M} = 1$, implying each of the N households has a single transaction, the variance of the NS1 estimator which would then depend on a srs of transactions with replacement is equivalent to the variance of the HH1 estimator that depends on a pps cluster sample of equivalent sample size selected with replacement.

4.3 Design Effects in Single Stage Sampling

Let

$$X'_{NS1} = \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}_j = \text{the unbiased NS estimator}$$

of X in single stage sampling, and

$$X'_{HH1} = \frac{M}{R_{HH}} \sum_{j=1}^{r_{HH}} \bar{X}_j = \text{the unbiased HH estimator of } X$$

in single stage sampling.

Define the single stage sampling total design effect of the NS1 estimator as the ratio of the variances of the NS1 and HH1 estimators of equivalent sample size conditional over all samples of n households.

$$\lambda(P) = \frac{\text{Var}(X'_{NS1})}{\text{Var}(X'_{HH1})} = \frac{\sigma_{NS1}^2(P)}{\bar{M} \sigma_{HH1}^2} \quad (18)$$

where $\lambda(P) < 1$ indicates that the NS1 estimator is more efficient than the HH1 estimator, and $\lambda(P) > 1$ indicates that the HH1 estimator is more efficient than the NS1 estimator.

We noted in (12) and (15) that $\sigma_{NS1}^2(P) = P \sigma_{NS1}^2 + P(1-P)(X/N^*)^2$, and in (16) that $\sigma_{HH1}^2 = \sigma_{NS1}^2(N^* = M)$. Making these substitutions in (18), the total design effect becomes

$$\lambda(P) = \text{def}_{NS1}^2 + (1-P) Z_{NS1}, \quad 0 < P \leq 1 \quad (19)$$

where

$$Z_{NS1} = \frac{P(X/N^*)^2}{\bar{M} \sigma_{NS1}^2(N^* = M)} \quad (20)$$

is the effect due to the N_o households without transactions, and

$$\text{def}_{NS1}^2 = \left[\frac{P \sigma_{NS1}^2}{\bar{M} \sigma_{HH1}^2} \right] = \left[\frac{P \sigma_{NS1}^2}{\bar{M} \sigma_{NS1}^2(N^* = M)} \right] \quad (21)$$

is effect due to the N^* households with transactions. In other words, def_{NS1}^2 is the design effect of *network sampling* a population of N^* household clusters containing one or more transactions, with equal probability and replacement, compared to *network sampling* a population of M transactions, of equivalent expected sample size, by srs and replacement. [The reader is referred to Kish (1982) for the definition of def^2].

The total design effect in (19) depends on def_{NS}^2 and Z_{NS1} and P , and the values of these parameters, as well as relationships between them, are likely to vary considerably between surveys, and between variables and population domains in the same surveys. Though, in theory, the NS1 estimator could be more efficient than HH1 estimator, in reality that outcome seems highly unlikely because cluster sampling is typically less efficient than srs. A necessary condition for the NS1 estimator to be as efficient or more efficient than the HH1 estimator is that $\text{def}_{NS1}^2 \leq 1 - (1-P)Z_{NS1}$, and this condition is unlikely to be met particularly if P is small, and if the within household transaction clustering is mostly due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments.

4.4 Comparing Efficiencies in Two-stage Sampling

In two stage sampling, the difference between the HH and NS second stage variance components for equivalent expected sample size of $nt\bar{M}$ transactions conditional over

all samples of n households, the second term on the right side of equation (11), reduces to

$$\begin{aligned} \frac{N}{nt} \sum_{j=1}^R \sigma_j^2 \left\{ (M_j - t) - \sum_{j=1}^N \frac{M_{ij}(M_j - t M_{ij})}{M_j} \right\} \\ = \frac{N}{n} \sum_{j=1}^R \frac{\rho_j}{M_j} \sigma_j^2 \end{aligned} \quad (22)$$

where $\rho_j/M_j = 1/M_j \sum_{i=1}^N M_{ij}(M_j - 1)$ is the difference between the HH and NS second stage finite population corrections for establishment j . If none of the N households have multiple transactions with establishment j , the HH and NS second stage variances of establishment j are equivalent and $\rho_j = 0$. Otherwise, $\rho_j > 0$ and second stage variance for establishment j is larger for the HH than the NS estimator. The value of ρ_j is maximum when establishment j has M_j transactions with a single household.

The second stage variance components of the HH and NS estimators are equivalent $\sum_{j=1}^R \rho_j = 0$, when, that is, none of the H households have multiple transactions with any of the R establishments. Of course, second stage variances are equivalent if transactions are selected with replacement or the within establishment variances, $\sigma_j^2 = 0$ ($j = 1, 2, \dots, R$). Except for these contingencies, however, the second stage variance is always larger for the HH estimator than for the NS estimator, and the magnitude of the difference depends on the extent of within household clustering of transactions with the same establishments, and the magnitudes of the within establishment variances.

If none of the N^* households have multiple transactions with the same establishments, the difference between the variances of the HH and NS estimators are equivalent in single stage and two stage establishment sample surveys. Otherwise, the difference between HH and NS variances is less in two stage than in single stage establishment sample surveys because whenever households have multiple transactions with the same establishments the second stage variance is greater for the HH estimator than for the NS estimator.

5. SUMMARY AND CONCLUDING REMARKS

The error model presented in this paper compares efficiencies of two estimators of the volume of transactions between establishments and populations in single-stage and two-stage establishment sample surveys. The Hansen-Hurwitz (HH) estimator depends on a stand-alone sampling frame that lists every establishment and the volume of its transactions with all households during a specified calendar period. The network sampling (NS) estimator depends on a population survey-generated frame that lists the households and their selection probabilities in a population sample survey, and for each household, lists the number of

its transactions with each distinct establishment during the specified calendar period.

Also, the NS and HH estimators depend on different establishment survey sample designs. In single-stage sampling, the HH estimator depends on a design in which establishments are the selection units and they are selected with pps with replacement, and the NS estimator depends on a design in which households are the selection units and they are selected with their selection probabilities in the population survey, which the error model assumes is srs with replacement. In two-stage sampling, transactions are the second stage sampling units of the HH and NS estimators. The HH estimator depends on fixed-size transaction samples that are selected by srs independently without replacement. The NS estimator depends on transaction sample sizes that are proportional to the number of transactions of each household with each establishment, and are selected independently by srs without replacement.

The NS and HH estimators are equally efficient, if and only if, every household in the entire population has one and only one transaction. Otherwise, neither the NS or the HH estimator is necessarily more efficient than the other. Nevertheless, it seems likely that the HH estimator will be more efficient than the NS estimator in single-stage establishment survey sampling, and perhaps substantially more efficient especially when large fractions of households do not have any transactions, and/or when the within household clustering of transactions among households with transactions is principally due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments. In two-stage sampling, the outcome is not as transparent as in single stage sampling because the second stage variance component is larger for the HH estimator than the NS estimator by an amount that depends on the extensiveness of within household clustering of transactions with the same establishments.

Arguably the foremost limitation of the error model presented in this paper is the presumption that the stand-alone and population survey-generated sampling frames are flawless in coverage and size measures. However, comparative costs of constructing and maintaining good quality stand-alone and population-generated establishment sampling frames are likely to vary greatly from survey to survey. Though the model seek to equalize the establishment survey costs based on each kind of sampling frames it ignores the differential costs of constructing and maintaining each kinds of frame.

Even in the absence of empirical data about the comparative costs of constructing and maintaining the frames, it is fair to say that the population survey-generated frame should be seriously considered as a potential design alternative whenever constructing and maintaining good quality stand-alone frames would be infeasible or exorbitantly expensive or time consuming, and/or when constructing and maintaining good quality population survey-generated

establishment sampling frames would be relatively inexpensive. For example, the population survey-generated frame would be a particularly attractive as a potential design alternative to the stand-alone frame when the stand-alone frame would be difficult to construct and maintain because it was undergoing rapid changing due to births, deaths, and establishment mergers, and the population survey-generated frame costs would be relatively small either because it could be constructed and maintained as a by-product of an ongoing population sample survey (Wunderlich 1992) and/or as a by-product of an ongoing program of matching transactions of households enumerated in a population survey with their establishment records (Cohen 1998).

Another limitation of the model is the unrealistic assumption that the population survey that generates the establishment sampling frame is based on a single stage sample design in which households are selected with equal probabilities and with replacement. In fact, population surveys are virtually always based on multistage sample designs in which households are selected without replacement in the final sampling stage. Typically, the srs assumption tends to significantly understate the variance of the NS estimator, and therefore would have the effect of exaggerating the relative efficiency of the NS estimator compared to the HH estimator. On the other hand, the household sampling with replacement assumption would have the opposite effects, but would be modest (Sirken 2001) compared to the srs assumption. The error model can be applied, however, to the other population survey sample designs that are not considered in this paper.

The error model presented in this paper identifies the critical parameters that determine the relative efficiency of establishment survey estimators depending on stand-alone and population survey-generated sampling frames. Values of these parameters will vary greatly between surveys and between variables and population domains in the same surveys. Unfortunately, empirical data are currently unavailable, and they are sorely needed to estimate the model's parameters under a broad range of survey conditions. Hopefully, this paper will stimulate interest in conducting establishment surveys that depend on population survey-generated establishment sampling frames, and will lead to improvements in designing establishment surveys that estimate the volume of transactions between establishments and populations.

ACKNOWLEDGMENTS

I thank the 2 referees and in particular an Assistant Editor for very helpful comments. The views expressed in this paper are solely those of the author and do not necessarily represent the official views or positions of the National Center for Health Statistics.

APPENDIX

When expressed as a function of P , the fraction of households with one or more transactions, the single stage population variance of the network sampling (NS) estimator of X

$$\sigma_{NS1}^2 = \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N} \right)^2$$

decomposes into 2 parts

$$\sigma_{NS1}^2(P) = P \sigma_{NS1}^2 + \sigma^2(P) E_{NS1}^2, \quad 0 < P \leq 1$$

where

$$P = \frac{N^*}{N},$$

$$\sigma_{NS1}^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2$$

is the truncated single stage population variance of the NS estimator exclusive of the $N_0 = N - N^*$ households without transactions with establishments,

$$\sigma^2(P) = P(1 - P)$$

is the variance of the binomial variable P , and

$$E_{NS1}^2 = (X/N^*)^2$$

is the expected value squared of the x -variate distributed over N^* households.

Proof

$$\begin{aligned} \sigma_{NS1}^2 &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{n} \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^{N^*} \left(\sum_{j=1}^R M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 + \frac{N_0}{N} \left(\frac{X}{N} \right)^2. \end{aligned} \quad (A.1)$$

Add and subtract X/N^* to the first term on the right side of (A.1).

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^{N^*} \left(\sum_{j=1}^R M_{ij} \bar{X}_j - \frac{X}{n} \right)^2 \\ &= \frac{P}{N^*} \sum_{i=1}^{N^*} \sum_{j \in A_i} \left(M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2 \\ &= P \sigma_{NS1}^2(P) + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2 \end{aligned} \quad (A.2)$$

Substitute (A.2) for the first term on the right side of (A.1).

$$\begin{aligned}\sigma_{NSI}^2(P) &= P \sigma_{NSI}^2 + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2 + (1-P) \left(\frac{X}{N} \right)^2 \\ &= P \sigma_{NSI}^2(P) + \sigma^2(P) E_{NSI}^2.\end{aligned}\quad (A.3)$$

where

$$\sigma^2(P) = P(1-P), \text{ and } E_{NSI}^2 = \left(\frac{X}{N^*} \right)^2.$$

REFERENCES

- COHEN, S.B. (1998). Sample design of the 1996 medical expenditure panel survey medical provider component. *Journal of Economic and Social Measurement*, 24, 25-53.
- DICKER, M., and SUNSHINE, J.H. (1987). Family use of health care, United States, 1980. *National Health Care Utilization and Expenditure Survey*. Report No. 10. DHHS Pub. 87-20210.
- JUDKINS, D., BERK, M., EDWARDS, S., MOHR, P., STEWART, K. and WAKSBERG, J. (1995). National Health Care Survey: List verses Network Sampling, Unpublished report. National Center for Health Statistics.
- JUDKINS, D., MARKER, D., WAKSBERG, J., BOTMAN, S. and MASSEY, J. (1999). National Health Interview Survey: Research for the 1995-2004 redesign. National Center for Health Statistics. *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2. 126, 76-89.
- KISH, L. (1982). Design effect. *Encyclopedia of the Statistical Sciences*. John Wiley & Sons, Inc. 2, 347-348.
- LEAVER, S., and VALLIANT, R. (1995). Statistical problems in estimating the U.S. consumer price index. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: John Wiley & Sons, Inc.
- MASSEY, L.T., MOORE, T.F., PARSONS, V. and TADRO, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2, 110.
- SIRKEN, M., and SHIMIZU, I. (1999). Population based establishment surveys: The Horvitz-Thompson estimator. *Survey Methodology*, 25, 187-91.
- SIRKEN, M., SHIMIZU, I. and JUDKINS, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 1, 470-473.
- SIRKEN, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics*. John Wiley & Sons, Inc. 4, 2977-2986.
- SIRKEN, M.G. (2001). The Hansen-Hurwitz estimator revisited: PPS sampling without replacement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. In print.
- THOMPSON, S. (1992). *Sampling*. New York: John Wiley & Sons, Inc. 117-118.
- WUNDERLICH, G.S. (Ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. Washington, DC: National Academy Press.

A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys

LOUIS-PAUL RIVEST¹

ABSTRACT

This paper suggests stratification algorithms that account for a discrepancy between the stratification variable and the study variable when planning a stratified survey design. Two models are proposed for the change between these two variables. One is a log-linear regression model; the other postulates that the study variable and the stratification variable coincide for most units, and that large discrepancies occur for some units. Then, the Lavallée and Hidiroglou (1988) stratification algorithm is modified to incorporate these models in the determination of the optimal sample sizes and of the optimal stratum boundaries for a stratified sampling design. An example illustrates the performance of the new stratification algorithm. A discussion of the numerical implementation of this algorithm is also presented.

KEY WORDS: Neyman allocation; Power allocation; Stratified random sampling.

1. INTRODUCTION

The construction of stratified sampling designs has a long history in the statistical sciences. Chapters 5 and 5A in Cochran (1977) review several techniques for splitting a population into strata. The construction of strata is a topic of current interest in the statistical literature. Recent contributions include Hedlin (2000) who revisits Ekman (1959) rule for stratification, and Dorfman and Valiant (2000) who compare model-based stratification with balanced sampling. Model based stratification, is discussed in Godfrey, Roshwalb, and Wright (1984) and in chapter 12 of Särndal, Swensson, and Wretman (1992).

In business surveys, populations have skewed distributions; a small number of units accounts for a large share of the total of the study variable. It is therefore appropriate to include all large units in the sample (Dalenius 1952; Glasser 1962). A good sampling design has one take-all stratum for big firms, where the units are all sampled, together with take-some strata for businesses of medium and small sizes. Typically the sampling fraction goes down with the size of the unit; small businesses get large sampling weights. The Lavallée and Hidiroglou (1988) stratification algorithm is often used to determine the stratum boundaries and the stratum sample sizes in this context (see for instance Slanta and Krenzke 1994, 1996). This algorithm uses a stratification variable, known for all the units of the population. It gives the stratum boundaries and the stratum sample sizes that minimize the total sample size required to achieve a target level of precision. It uses an iterative procedure, due to Sethi (1963), to determine the optimal stratum boundaries. The Lavallée and Hidiroglou algorithm does not account for a difference between the stratification and the survey variables. As time goes by, this difference increases and the sampling design provided by

the Lavallée and Hidiroglou algorithm may fail to meet the precision criterion.

Stratification in situations where the survey variable and the stratification variable differ is considered in Dalenius and Gurney (1951), see also Cochran (1977, chapter 5A). Many authors have studied approximate formulae for determining stratum boundaries, and for evaluating the gain in precision resulting from stratification on an auxiliary variable. Some relevant contributions are Serfling (1968), Singh and Sukatme (1969), Singh (1971), Singh and Parkash (1975), Anderson, Kish and Cornell (1976), Oslo (1976), Wang and Aggarwal (1984) and Yavada and Singh (1984). Hidiroglou and Srinath (1993) and Hidiroglou (1994) suggest techniques to update stratum boundaries using a new stratification variable. However these papers do not explicitly provide stratification algorithms accounting for the discrepancy between the stratification variable and the survey variable. This paper fills this gap by constructing generalizations of the Lavallée and Hidiroglou (1988) algorithm that express the difference between these two variables in terms of a statistical model.

A brief review of stratified sampling and of sample allocation methods is first given. Models for the difference between stratification and survey variables are then proposed. The implementation of Sethi's algorithm, when the stratification and the survey variable differ, is then presented. Numerical illustrations are provided.

2. A REVIEW OF STRATIFIED RANDOM SAMPLING

Some of the standard notation of stratified random sampling that will be used in this paper is

L = the number of strata;

¹ Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4.

$W_h = N_h/N$ is for $h = 1, \dots, L$ the relative weight of stratum h , N_h is the size of stratum h , and $N = \sum N_h$ is the total population size;

n_h is for $h = 1, \dots, L$ the sample size in stratum h and $f_h = n_h/N_h$ is the sampling fraction;

\bar{Y}_h and \bar{y}_h are the population and sample means of Y within stratum h ;

S_{yh} is the population standard deviation of Y within stratum h .

In this paper the strata are constructed using X , a stratification variable. Stratum h consists of all units with an X -value in the interval $(b_{h-1}, b_h]$, where $-\infty = b_0 < b_1 < \dots < b_{L-1} < b_L = \infty$ are the stratum boundaries.

The survey estimator for \bar{Y} can be expressed as $\bar{y}_{st} = \sum W_h \bar{y}_h$; its variance is given by:

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2. \quad (2.1)$$

In business surveys, all the big firms are sampled; we choose stratum L as the take-all stratum so that $n_L = N_L$. For $h < L$, n_h , the sample size in take-some stratum h , can be written as $(n - N_L)a_h$ where n is the total sample size and a_h depends on the allocation rule. The two allocation rules that are considered in this paper are

- The power allocation rule

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^{L-1} (W_k \bar{Y}_k)^p} \quad (2.2)$$

where p is a positive number in $(0, 1]$;

- The Neyman allocation rule

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}}. \quad (2.3)$$

Solving (2.1) for n leads to

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2 / a_h}{\text{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2 / N}. \quad (2.4)$$

The optimal stratum boundaries are the values of b_1, \dots, b_{L-1} that minimize n subject to a requirement on the precision of \bar{y}_{st} such as $\text{Var}(\bar{y}_{st}) = \bar{Y}^2 c^2$ where c is the target coefficient of variation (CV). The range $c = 1\%$ to 10% is often used for business surveys.

3. SOME MODELS FOR THE DISCREPANCY BETWEEN THE STRATIFICATION AND THE SURVEY VARIABLE

In this section $\{x_i, i = 1, \dots, N\}$ denotes the known stratification variable for the N units in the population. Many stratification algorithms, including Lavallée and Hidirolglou, suppose that $\{x_i, i = 1, \dots, N\}$ also represents the values of the study variable. This section suggests statistical models to account for a difference between these two variables.

For the sequel, it is convenient to look at X and Y as continuous random variables and to let $f(x)$, $x \in \mathbb{R}$ denote the density of X . The data $\{x_i, i = 1, \dots, N\}$ can be viewed as N independent realizations of the random variable X . Since stratum h consists of the population units with an X -value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ and $\text{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h , for $h = 1, \dots, L-1$. Three models for the difference between X and Y are next given along with their conditional means and variances for Y .

3.1 A Log-linear Model

The first model considers that $\log(Y) = \alpha + \beta_{\log} \log(X) + \epsilon$, where ϵ is a normal random variable with mean 0 and variance σ_{\log}^2 , which is independent from X , and α and β_{\log} are parameters to be determined. When $\alpha = 0$, $\beta_{\log} = 1$ and $\sigma_{\log}^2 = 0$, one has $X = Y$; the survey and the stratification variables are the same. In general, $Y = e^\alpha X^{\beta_{\log}} e^\epsilon$. The conditional moments of Y can be evaluated using the basic properties of the lognormal distribution (see Johnson and Kotz 1970), that is

$$E(e^\epsilon) = e^{\sigma_{\log}^2/2} \text{ and } \text{Var}(e^\epsilon) = e^{\sigma_{\log}^2}(e^{\sigma_{\log}^2} - 1).$$

One has

$$E(Y|b_h \geq X > b_{h-1}) = \exp(\alpha + \sigma_{\log}^2/2) E(X^{\beta_{\log}}|b_h \geq X > b_{h-1})$$

while $\text{Var}(Y|b_h \geq X > b_{h-1})$ is equal to

$$\begin{aligned} & \text{Var}(E(Y|X)|b_h \geq X > b_{h-1}) + E(\text{Var}(Y|X)|b_h \geq X > b_{h-1}) \\ &= \exp(2\alpha + \sigma_{\log}^2) \{ \text{Var}(X^{\beta_{\log}}|b_h \geq X > b_{h-1}) \\ & \quad + (e^{\sigma_{\log}^2} - 1) E(X^{2\beta_{\log}}|b_h \geq X > b_{h-1}) \} \\ &= \exp(2\alpha + \sigma_{\log}^2) \{ e^{\sigma_{\log}^2} E(X^{2\beta_{\log}}|b_h \geq X > b_{h-1}) \\ & \quad - E(X^{\beta_{\log}}|b_h \geq X > b_{h-1})^2 \}. \end{aligned}$$

The parameter values β_{\log} and σ_{\log} can sometimes be calculated from historical data. Simple ad hoc values are $\beta_{\log} = 1$ and $\sigma_{\log}^2 = (1 - \rho^2) \text{Var}(\log(X))$. Here ρ is the assumed correlation between $\log(X)$ and $\log(Y)$. It can be set equal to predetermined values such as 0.95 or 0.99.

3.2 A Linear Model

In the survey sampling literature, the discrepancy between Y and X is often modeled with a heteroscedastic linear model,

$$Y = \beta_{\text{lin}} X + \varepsilon, \tag{3.5}$$

where the conditional distribution of ε , given X , has mean 0 and variance $\sigma_{\text{lin}}^2 X^\gamma$, for some non negative parameter γ . Straightforward calculations lead to $E(Y|b_h \geq X > b_{h-1}) = \beta_{\text{lin}} E(X|b_h \geq X > b_{h-1})$ while $\text{Var}(Y|b_h \geq X > b_{h-1}) = \beta_{\text{lin}}^2 \{ \text{Var}(X|b_h \geq X > b_{h-1}) + (\sigma_{\text{lin}}/\beta_{\text{lin}})^2 E(X^\gamma|b_h \geq X > b_{h-1}) \}$.

For an arbitrary $\gamma \geq 0$, the conditional variance of Y depends on three conditional moments of X . The generalization of Sethi's algorithm presented in section 5 does not work in this situation. Note however that when $\gamma = 2$, the conditional mean and variance of Y are proportional to those for the log-linear model with

$$\beta_{\log} = 1 \text{ and } \sigma_{\log}^2 = \log(1 + (\sigma_{\text{lin}}/\beta_{\text{lin}})^2); \tag{3.6}$$

the proportionality factors are $\exp(\alpha + \sigma_{\log}^2/2)/\beta_{\text{lin}}$ and $\exp(2\alpha + \sigma_{\log}^2)/\beta_{\text{lin}}^2$ for the conditional expectations and the conditional variances respectively. Thus the two models for the discrepancy between the stratification and the survey variable, either the log-linear model of section 3.1 or the linear model (3.5) with parameter $\gamma = 2$, lead, in section 5, to the same stratified design provided that (3.6) holds. In the later sections, the log-linear model is used to represent the change between X and Y . It should give good results when the true relationship between Y and X is modeled by (3.5) with $\gamma \approx 2$. When model (3.5) is assumed to hold with a smaller value of γ , the algorithm of section 5 can still be implemented when γ is set to either 0 or 1. This is however not pursued in this paper.

3.3 A Random Replacement Model

This model assumes that the stratification variable is equal to the survey variable, *i.e.*, $X = Y$, for most units. There is however a small probability ε that a unit changed drastically; its Y value then has $f(x)$ as density and is distributed independently of its X value. This is the approach used in Rivest (1999) to model the occurrence of stratum jumpers for which X is not representative of Y . More formally, this can be written as,

$$Y = \begin{cases} X & \text{with probability } 1 - \varepsilon \\ X_{\text{new}} & \text{with probability } \varepsilon \end{cases},$$

where X_{new} represents a random variable with density $f(x)$ distributed independently of X . The conditional mean for Y under this model is given by

$$E(Y|b_h \geq X > b_{h-1}) = (1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X),$$

while its conditional variance is equal to

$$\begin{aligned} \text{Var}(Y|b_h \geq X > b_{h-1}) \\ &= (1 - \varepsilon) E(X^2|b_h \geq X > b_{h-1}) + \varepsilon E(X^2) \\ &\quad - \{ (1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X) \}^2. \end{aligned}$$

4. AN EXAMPLE

Before addressing the technical details underlying the construction of the algorithms, it is convenient to look at an example. Consider the MU284 population of Särndal, Swensson and Wretman (1992), presenting data on 284 Swedish municipalities.

To build a stratified design for estimating the average of RMT85, the revenues from the 1985 municipal taxation, REV84, the real estate value according to 1984 assessment, is used as a stratification variable. One takes $L = 5$ and set the target CV at 5%. Two stratified designs obtained with the Lavallée and Hidiroglou algorithm are given in Table 1, for the power allocation with $p = 0.7$ and the Neyman allocation. Both have $n = 19$. When applied on survey variable RMT85, these two designs give estimators of total revenue with coefficients of variation of 8.3% and 7.3% respectively. Failing to account for a change between the survey and the stratification variables yields estimators that are more variable than expected.

Table 1
Stratified designs obtained with the Lavallée and Hidiroglou algorithm for the MU284 population using REV84 as stratification variable and a target CV of 5%

Power allocation with $p = 0.7$							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,251	874	56,250	86	1	0.01	19
stratum 2	2,352	1,696	100,898	82	2	0.02	19
stratum 3	4,603	3,114	351,547	65	3	0.05	19
stratum 4	10,606	6,442	2,027,436	41	3	0.07	19
stratum 5	59,878	19,631	275,502,518	10	10	1	19
Neyman allocation							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,273	878	57,260	87	2	0.02	19
stratum 2	2,336	1,701	99,688	81	2	0.02	19
stratum 3	4,619	3,114	351,547	65	3	0.05	19
stratum 4	11,776	6,921	3,724,610	46	7	0.15	19
stratum 5	59,878	28,418	426,851,844	5	5	1	19

To model the discrepancy between REV84 and RMT85, we use the log-linear model of section 3.1. There are outliers in the linear regression of $\log(\text{RMT85})$ on $\log(\text{REV84})$; they make the least squares estimates of β_{\log} and σ_{\log} unrepresentative of the relationship between the two variables. Robust estimates obtained with the Splus function `lmRobMM` are used instead. They are given by $\hat{\beta}_{\log} = 1.1$ and $\hat{\sigma}_{\log} = 0.2116$. Table 2 gives the stratified designs obtained with the generalized Lavallée and Hidirolglou algorithm for two allocation rules. They both give estimators of the total of RMT85 having a CV of 5.7%. This CV is still larger than 5%. Since there are outliers in the log-linear regression, the assumption of normal errors made in section 3.1 is not met. This might explain the failure to reach the target CV exactly. The increase in sample size for $n = 19$ to $n = 28$ is noteworthy! For both allocation methods the design obtained using the log-linear model has smaller take-all strata than Lavallée and Hidirolglou.

Table 2

Stratified designs obtained with the generalized Lavallée and Hidirolglou algorithm for the MU284 population using REV84 as stratification variable, a log-linear with $\beta_{\log} = 1.1$ and $\sigma_{\log} = 0.2116$ for the discrepancy between REV84 and RMT85, and a target CV of 5%

Log-linear model stratification algorithm with power allocation with $p = 0.7$							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,558	1,023	97,245	121	4	0.03	28
stratum 2	3,031	2,219	168,204	81	5	0.06	28
stratum 3	5,706	4,022	464,471	44	6	0.14	28
stratum 4	11,107	7,602	2,659,061	32	7	0.22	28
stratum 5	59,878	25,536	39,131,413	6	6	1	28
Log-linear model stratification algorithm with Neyman allocation							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1,582	1,023	97,245	121	4	0.03	28
stratum 2	3,040	2,219	168,204	81	5	0.06	28
stratum 3	5,608	4,022	464,471	44	5	0.11	28
stratum 4	11,476	7,709	2,952,313	33	9	0.27	28
stratum 5	59,878	28,418	426,851,844	5	5	1	28

An alternative to the generalized Lavallée and Hidirolglou algorithm for the construction of stratified designs is to use their original algorithm with a smaller target CV. This increases the sample size thereby reducing the variance of the estimator of the total of the survey variable. When constructing a design for RMT85 using REV84 as a stratification variable, the standard Lavallée and Hidirolglou algorithm with power allocation rule ($p = 0.7$) and a target CV of 3.6%, yields a stratified design with $n = 28$. This design has the same sample size as those presented in Table 2. The CV of the estimator of the total RMT85 is 5.7%, the

same as the CVs obtained with the designs of Table 2. The main difference between these designs is the size of the take-all stratum. The design constructed with the Lavallée and Hidirolglou algorithm has a take-all stratum of size $N_5 = 13$ as compared to $N_5 = 5$ and $N_5 = 6$ for the designs of Table 2. Allowing the stratification and the survey variables to differ appears to reduce the relative importance of the take-all stratum in the sampling design. Further investigations are needed to ascertain this hypothesis.

The stratification algorithm for the random replacement model of section 3.3 (with Neyman allocation) was also applied to REV84. Assuming changes in 2% of the units ($\varepsilon = 0.02$), the generalized Lavallée and Hidirolglou algorithm yields a stratified design with $n = 37$ sample units; the resulting estimator of total RMT85 has a CV of 5.5%. An interesting property of this stratified design is that the smallest sampling fraction is $\min_h f_h = 9.3\%$; it is much larger than $\min_h f_h$ for the designs of Tables 1 and 2. Despite the presence of outliers, the random replacement model does not describe the changes between REV84 and RMT85 as well as the log-linear model. This explains why a larger sample size, 37 instead of 28, is needed to get an estimator with a variance comparable to that obtained with the stratification based on a log-linear model.

5. A METHOD FOR CONSTRUCTING STRATIFICATION ALGORITHMS

The aim of a stratification algorithm is to determine the optimal stratum boundaries and sample sizes for sampling Y using the known values $\{x_i; i = 1, \dots, N\}$ of variable X for all the units in the population. A model, such as those given in section 3, characterizes the relationship between X and Y . This section extends the stratification algorithm of Lavallée and Hidirolglou (1988) to situations where X and Y differ. It uses the log-linear model of section 3.1 to account for the differences between Y and X . Modifications to handle the random replacement model are easily carried out (see Rivest 1999).

5.1 A Generalization of Sethi's (1963) Stratification Method

It is convenient to consider an infinite population analogue to equation (2.4) for n . Since the random variable X has a density $f(x)$, the first two conditional moments of Y given that $b_{h-1} < X \leq b_h$ can be written in terms of

$$W_h = \int_{b_{h-1}}^{b_h} f(x) dx, \quad \Phi_h = \int_{b_{h-1}}^{b_h} \alpha^\beta f(x) dx,$$

$$\text{and } \Psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x) dx,$$

where β is the slope of the log-linear model given in section 3.1 (in this section β and σ represent parameters of the log-linear model of section 3.1, since there is no risk of

confusion the subscript log is not used anymore). For stratification purposes, it is useful to rewrite (2.4) in terms of the conditional means and variances for Y ,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \text{Var}(Y|b_h \geq X > b_{h-1})/a_{h,X}}{\bar{Y}^2 C^2 + \sum_{h=1}^{L-1} W_h \text{Var}(Y|b_h \geq X > b_{h-1})/N}, \quad (5.7)$$

where $a_{h,X}$ denotes the allocation rule written in terms of the known X . For instance, under power allocation,

$$a_{h,X} = \frac{\{W_h E(Y|b_h \geq X > b_{h-1})\}^p}{\sum_{k=1}^{L-1} \{W_k E(Y|b_k \geq X > b_{k-1})\}^p},$$

for $h = 1, \dots, L-1$. Given a model for the relationship between Y and X , $\text{Var}(Y|b_h \geq X > b_{h-1})$ and $E(Y|b_h \geq X > b_{h-1})$ can be written in terms of W_h , ϕ_h , and ψ_h . Thus, the partial derivatives of n with respect to b_h can be evaluated, for $h < L-1$, using the chain rule,

$$\begin{aligned} \frac{\partial n}{\partial b_h} &= \frac{\partial n}{\partial W_h} \frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \phi_h} \frac{\partial \phi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h} \frac{\partial \psi_h}{\partial b_h} \\ &+ \frac{\partial n}{\partial W_{h+1}} \frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \phi_{h+1}} \frac{\partial \phi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}} \frac{\partial \psi_{h+1}}{\partial b_h} \end{aligned}$$

Observe that

$$\begin{aligned} \frac{\partial W_h}{\partial b_h} &= -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h) \\ \frac{\partial \phi_h}{\partial b_h} &= -\frac{\partial \phi_{h+1}}{\partial b_h} = b_h^\beta f(b_h) \\ \frac{\partial \psi_h}{\partial b_h} &= -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} f(b_h) \end{aligned}$$

This leads to the following result, for $h < L-1$,

$$\frac{\partial n}{\partial b_h} = f(b_h)$$

$$\left\{ \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) + \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right) b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) b_h^{2\beta} \right\}.$$

Similarly,

$$\frac{\partial n}{\partial b_{L-1}} = f(b_{L-1}) \left\{ -N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \phi_{L-1}} b_{L-1}^\beta + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^{2\beta} \right\}.$$

The Sethi's (1963) algorithm is used to solve $\partial n / \partial b_h = 0$. It considers that the partial derivatives are proportional to quadratic functions in b_h^β . The updated value for b_h^β is given by the largest root of the corresponding quadratic function. When $h < L-1$, this gives

$$\begin{aligned} b_h^{\beta \text{ new}} &= \frac{-\left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right) / \left\{ 2 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \right\}}{\left\{ \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right)^2 - 4 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) \right\}^{1/2}} \\ &+ \frac{\left\{ 2 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \right\}}{\left\{ \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right)^2 - 4 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) \right\}^{1/2}}, \end{aligned}$$

while for $h = L-1$ we have

$$b_{L-1}^{\beta \text{ new}} = \frac{-\frac{\partial n}{\partial \phi_{L-1}} + \left\{ \left(\frac{\partial n}{\partial \phi_{L-1}} \right)^2 - 4 \frac{\partial n}{\partial \psi_{L-1}} \left(\frac{\partial n}{\partial W_{L-1}} - N \right) \right\}^{1/2}}{\left(2 \frac{\partial n}{\partial \psi_{L-1}} \right)}$$

The partial derivatives of n with respect to W_h , ϕ_h , and ψ_h depend on moments of order 0, 1, and 2 of x^β within stratum h . These moments are evaluated using the N x -values in the population. For instance,

$$\phi_h = \frac{1}{N} \sum_{i: b_{h-1} < x_i \leq b_h} x_i^\beta.$$

Applications of this general method are provided next.

When using Sethi's algorithm, one typically has $L \geq 3$. Note however that it also works when $L = 2$. In this case, the algorithm is searching for the boundary between a take-all and a take-some stratum. Successive evaluations of $b_{L-1}^{\beta \text{ new}}$ presented above yield an optimal boundary. When one assumes that the stratification and the study variable coincide, i.e., $X = Y$, this boundary is nearly identical to that obtained with the algorithm presented in Hidirolglou (1986).

5.2 An Algorithm for Power Allocation

For the log-linear model of section 3.1, the conditional expectation is $E(Y|b_h \geq X > b_{h-1}) = C \phi_h / W_h$ while the conditional variance is

$$\text{Var}(Y|b_h \geq X > b_{h-1}) = C^2 \{ e^{\sigma^2} \psi_h / W_h - (\phi_h / W_h)^2 \},$$

where $C = \exp(\alpha + \sigma^2/2)$. Under the power allocation rule, $a_{h,X} = \phi_h^p / \sum_{h=1}^{L-1} \phi_h^p$, and formula (5.7) for n becomes

$$n = NW_L + \frac{\sum_{h=1}^{L-1} \phi_h^p \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h / W_h - \phi_h^2 / W_h) / \phi_h^p}{\left(\sum x_i^\beta / N \right)^2 C^2 + \sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h - \phi_h^2 / W_h) / N}.$$

The partial derivatives needed to implement the stratification algorithm are easily calculated; for $h \leq L-1$,

$$\frac{\partial n}{\partial W_h} = \frac{A e^{\sigma^2 \psi_h / \varphi_h^p}}{F} - \frac{AB(\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{A \{-pe(\sigma^2 W_h \psi_h - \varphi_h^2) / \varphi_h^{p+1} - 2 / \varphi_h^{p-1}\} + p \varphi_h^{p-1} B}{F} + 2 \frac{AB \varphi_h / (n W_h)}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = e^{\sigma^2 A W_h / \varphi_h^p} \frac{A}{F} - e^{\sigma^2} \frac{AB / N}{F^2},$$

where

$$A = \sum_{h=1}^{L-1} \varphi_h^p, B = \sum_{h=1}^{L-1} \left(e^{\sigma^2 W_h \psi_h - \varphi_h^2} \right) / \varphi_h^p,$$

and

$$F = \left(\sum x_i^\beta / N \right)^2 c^2 + \sum_{h=1}^{L-1} \left(e^{\sigma^2 \psi_h - \varphi_h^2 / W_h} \right) / N.$$

5.3 An algorithm for Neyman allocation

Under Neyman allocation, allocation rule (2.3) written in terms of W_h , φ_h , and ψ_h is

$$a_{h,X} = \frac{\left\{ e^{\sigma^2 \psi_h W_h - \varphi_h^2} \right\}^{1/2}}{\sum_{h=1}^{L-1} \left\{ e^{\sigma^2 \psi_h W_h - \varphi_h^2} \right\}^{1/2}}$$

and the formula for n is

$$n = NW_L + \frac{\left\{ \sum_{h=1}^{L-1} (e^{\sigma^2 \psi_h W_h - \varphi_h^2})^{1/2} \right\}^2}{\left(\sum x_i^\beta / N \right)^2 c^2 + \sum_{h=1}^{L-1} (e^{\sigma^2 \psi_h - \varphi_h^2 / W_h}) / N}.$$

The partial derivatives needed to implement Sethi's (1963) iterative algorithm are,

$$\frac{\partial n}{\partial W_h} = \frac{A e^{\sigma^2 \psi_h / (e^{\sigma^2 \psi_h W_h - \varphi_h^2})^{1/2}}}{F} - \frac{A^2 (\varphi_h / W_h)^2 / N}{F^2}$$

$$\frac{\partial n}{\partial \varphi_h} = \frac{-2A \varphi_h / \{e^{\sigma^2 W_h \psi_h - \varphi_h^2}\}^{1/2}}{F} + \frac{2A^2 \varphi_h / (W_h N)}{F^2}$$

$$\frac{\partial n}{\partial \psi_h} = \frac{e^{\sigma^2 A W_h / \{e^{\sigma^2 W_h \psi_h - \varphi_h^2}\}^{1/2}}}{F} - e^{\sigma^2} \frac{A^2 / N}{F^2},$$

where

$$A = \sum_{h=1}^{L-1} \left(e^{\sigma^2 \psi_h W_h - \varphi_h^2} \right)^{1/2},$$

and

$$F = \left(\sum x_i^\beta / N \right)^2 c^2 + \sum_{h=1}^{L-1} \left(e^{\sigma^2 \psi_h - \varphi_h^2 / W_h} \right) / N.$$

6. NUMERICAL CONSIDERATIONS

Slanta and Krenzke (1994, 1996) encountered numerical difficulties when using the Lavallée and Hidirolglou algorithm with Neyman allocation: convergence was slow and sometimes the algorithm did not converge to the true minimum value for n . Indeed Schneeberger (1979) and Slanta and Krenzke (1994) showed that, for a particular bimodal population, the problem has a saddle; that is the partial derivatives are all null at boundaries b_h which do not give a true minimum for n .

When using the algorithms constructed in this paper, we also experienced the numerical difficulties alluded to in Slanta and Krenzke (1994). The algorithms constructed under power allocation were generally more stable than those using Neyman allocation; numerical difficulties were more frequent when the number L of strata was large. Furthermore, as the distribution for Y moved away from that of X , i.e., as σ^2 increases, non convergence of the algorithm and failure to reach the global minimum for n were more frequent. In these situations, the stratification algorithm's starting values were of paramount importance. For instance, in Table 2, the design accounting for changes between Y and X obtained under Neyman allocation depends heavily on the starting values. The one presented in Table 2 uses the boundaries presented in Table 2 for the power allocation as starting values. Starting the algorithm with the boundaries obtained in Table 1 for the Lavallée Hidirolglou algorithm with Neyman allocation yields a different sampling design having $n = 29$.

A good numerical strategy is to run the stratification algorithm for several intermediate designs to get to a final sampling design, with the stratum boundaries obtained at one step used as starting values for the algorithm at the next step. The log-linear algorithm is always run in two steps; first run the Lavallée and Hidirolglou algorithm, setting $\sigma = 0$, and use these boundaries as starting value for the algorithm with a non null σ . Also use as starting value for Neyman allocation the corresponding boundaries found under power allocation with a p value around 0.7.

7. CONCLUSION

This paper has proposed generalizations of the Lavallée and Hidioglou stratification algorithm that account for a difference between the stratification and the survey variables. Two statistical models have been introduced for this purpose. The new class of algorithms uses the Chain Rule to derive partial derivatives and Sethi's (1963) technique to find the optimal stratum boundaries.

The log-linear model stratification algorithm proposed in this paper was used successfully in several surveys designed at the Statistical Consulting Unit of Université Laval. For estimating total maple syrup production in a year, the number of sap producing maples for a producer was a convenient size variable. Historical data was used to estimate the parameters of the log-linear model linking sap producing maples and production volume. Another example is the estimation of the total maintenance deficit of hospital buildings in Quebec. The value of each building was the known stratification variable. The maintenance deficit was estimated to be in the range (20%, 40%) by experts. Solving $4\sigma_{\log} = \log(40\%) - \log(20\%)$ gives $\sigma_{\log} = \log(2)/4 = 0.17$ as a possible parameter value for the log-linear model of section 3.1. In these two examples accounting for changes between the stratification and the survey variables increased the sample size n by a fair percentage and yielded survey estimators whose estimated CVs were close to the target CVs.

Two SAS IML functions implementing the algorithm presented in this paper, for power and Neyman allocation, are available on the author's website at <http://www.mat.ulaval.ca/pages/lpr/>. They allow user specified starting values for the stratum boundaries; they can be used to implement the numerical strategies presented in section 6.

ACKNOWLEDGMENTS

I am grateful to Nathalie Vandal and to Gaétan Daigle for constructing SAS IML programs for the stratification algorithms used in the paper. The constructive comments of the associate editor and of the referee are gratefully acknowledged.

REFERENCES

- ANDERSON, D.W., KISH, L. and CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*. 71, 887-892.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition. New York: John Wiley & Sons, Inc.
- DALENIUS, T. (1952). The Problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*. 35, 61-70.
- DALENIUS, T., and GURNEY, M. (1951). The Problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*. 34, 133-148.
- DORFMAN, A.H., and VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*. 16, 139-154.
- ECKMAN, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*. 30, 219-229.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*. 30, 28-32.
- GODFREY, J., ROSHWALB, A. and WRIGHT, R.L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*. 2, 1-9.
- HEDLIN, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*. 16, 15-29.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*. 40, 27-31.
- HIDIROGLOU, M. (1994). Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 153-162.
- HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*. 11, 397-405.
- JOHNSON, N.L., and KOTZ, S. (1970). *Continuous Univariate Distribution-1*. New York: John Wiley & Sons, Inc.
- LAVALLÉE, P., and HIDIROGLOU, M. (1988). On the stratification of skewed populations. *Survey Methodology*. 14, 33-43.
- OSLO, I.T. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*. 23, 15-25.
- RIVEST, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 64-72.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- SCHNEEBERGER, H. (1979). Saddle points of the variance of the sample mean in stratified sampling. *Sankhyā: The Indian Journal of Statistics, Series C*. 41, 92-96.
- SERFLING, R.J. (1968). Approximate optimal stratification. *Journal of the American Statistical Association*. 63, 1298-1309.
- SETHI, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*. 5, 20-33.
- SINGH, R.J. (1971). Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*. 66, 829-834.
- SINGH, R., and PARKASH, D. (1975). Optimal stratification for equal allocation. *Annals of the Institute of Statistical Mathematics*. 27, 273-280.
- SINGH, R., and SUKATME, B.V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*. 21, 515-528.

- SLANTA, J., and KRENZKE, T. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 693-698.
- SLANTA, J., and KRENZKE, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Survey Methodology*. 22, 65-75.
- WANG, M.C., and AGGARWAL, V. (1984). Stratification under a particular Pareto distribution. *Communications in Statistics, Part A – Theory and Methods*. 13, 711-735.
- YAVADA, S., and SINGH, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics, Part A – Theory and Methods*. 13, 2793-2806.

Multi-way Stratification by Linear Programming Made Practical

WILSON LU and RANDY R. SITTER¹

ABSTRACT

Sitter and Skinner (1994) present a method which applies linear programming to designing surveys with multi-way stratification, primarily in situations where the desired sample size is less than or only slightly larger than the total number of stratification cells. The idea in their approach is simple, easily understood and easy to apply. However, the main practical constraint of their approach is that it rapidly becomes expensive in terms of magnitude of computation as the number of cells in the multi-way stratification increases, to the extent that it cannot be used in most realistic situations. In this article, we extend this linear programming approach and develop methods to reduce the amount of computation so that very large problems become feasible.

KEY WORDS: PPS sampling; Proportional allocation; Random grouping; Survey sampling.

1. INTRODUCTION

In many practical survey situations, there are multiple stratifying variables available and thus the designer has the option of defining strata as cells formed as cross-classified categories of these variables. For examples, see Engle, Marsden and Pollock (1971), Hess, Riedel and Fitzpatrick (1976), Vihma (1981) and Skinner, Holmes and Holt (1994). This multi-way stratification often leads to situations where the desired sample size is less than or only slightly larger than the total number of stratification cells (particularly common when choosing primary sampling units (psu's) in stratified multi-stage designs) and hence conventional methods of sample allocation to strata may not be applicable.

An illustration, based on a hypothetical example of Bryant, Hartley and Jessen (1960), is given in Table 1. Communities (psu's) are classified by two stratifying factors, type and region, with three and five categories respectively. The desired sample size of $n = 10$ is less than the total number of cells, 15. This example also illustrates a related problem. The entries in Table 1 are the expected counts under proportional stratification, *i.e.*, the strata sample sizes are proportional to the population strata sizes. Under the sample size restrictions, the expected cell sample counts will not generally be integers. In cases with very small expected counts, rounding to integers will not lead to good choices while causing a serious violation of the property of proportional allocation. Non-integer margin totals are also typical and can cause their own difficulties. Goodman and Kish (1950) was the first to address this problem under the name of controlled selection, where they propose a sampling selection procedure which can be classified as random systematic sampling (see Hess, Riedel and Fitzpatrick 1976; Waterton 1983). Bryant *et al.* (1960) presented a very simple method to randomly assign sample

sizes for each cell in two-way stratification and gave two estimators based on that sampling scheme. However, since the expected cell sample sizes didn't include information of proportion of each cell (*i.e.*, the method is not a proper controlled selection technique, as only the probabilities of the marginal distributions are respected), these estimators may not have satisfactory MSE properties (see Sitter and Skinner 1994). Jessen (1970) points out that a further limitation of the method of Bryant *et al.* (1960) is that it is not possible to constrain specified cell sizes to be zero, which may be desired in some situations (see related methods under the label "lattice sampling", *e.g.* Jessen 1973, 1975). He proposes two methods for both two-way and three-way stratification but both methods are fairly complicated to implement and, as noted by Causey, Cox and Ernst (1985), may not lead to a solution. Inspired by the idea of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples (see also Lahiri and Mukerjee 2000), Sitter and Skinner (1994) proposed a linear programming approach which attempts to take advantage of the power of modern computing. This linear programming technique is simple in conception, is flexible to different situations, always has a solution and has better properties of the MSE. Its main practical constraint is that it becomes computationally intensive as the number of cells in the multi-way stratification increases, quickly to the point of infeasibility. In this paper we will present a simple method which will allow the linear programming technique to handle much larger problems. In section 2 we describe the linear programming method of Sitter and Skinner (1994) to introduce notation and briefly discuss its numerical limitations. In section 3.1, we first discuss some simple strategies to reduce the computational intensity of the method as motivation for the eventual proposal. In sections 3.2 and 3.3 we discuss the proposed method assuming integer margins

¹ Wilson Lu, Doctoral Student, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

and give some examples with from 80 to 300 stratification cells to illustrate the ability of the new methodology to handle large problems. In section 3.4, we describe the simple extension of the method to non-integer margins and illustrate by applying the method to a real example from the occupational health literature (Vihma 1981).

Table 1
Example from Bryant *et al.* (1960). Expected Sample Cell Counts Under Proportional Stratification ($n = 10$)

Region	Type of Community			Total
	Urban	Rural	Metropolitan	
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

2. THE LINEAR PROGRAMMING TECHNIQUE

2.1 The Basic Ideas

We introduce the linear programming method of Sitter and Skinner (1994) by considering the simplest kind of two-way stratification. Suppose that N units of a finite population are arranged in a two-way classification in R rows formed by categories of one variable and C columns by categories of another. Let N_{ij} denote the number of population units in the i -th row and the j -th column (*i.e.*, in the ij -th cell) of the two-way table and $P_{ij} = N_{ij}/N$ denote the proportion of the total population in the ij -th cell. Let \bar{Y} denote the mean value of a survey characteristic y for the population and \bar{Y}_{ij} denote the mean value of y for the ij -th cell.

The sample is selected as follows:

- Sample sizes n_{ij} are randomly determined for each cell according to a pre-specified procedure. Letting s denote the $R \times C$ array $(n_{ij}, i = 1, \dots, R, j = 1, \dots, C)$, this procedure assigns a probability $p(s)$ to each s in the set S of possible such arrays and selects a single array, s , from S . We denote the dependence of n_{ij} on s by writing $n_{ij}(s)$.
- A simple random sample of $n_{ij}(s)$ units is then selected from the ij -th cell and the values of y obtained.

Restrict attention to designs of fixed sample size $n > 0$, that is, restrict to arrays $s \in S_n$ such that $\sum_{i=1}^R \sum_{j=1}^C n_{ij}(s) = n$. We would also like to restrict attention to proportionate stratification so that

$$\sum_{s \in S_n} n_{ij}(s) p(s) = n P_{ij} \quad \text{for } i = 1, \dots, R, j = 1, \dots, C, \quad (1)$$

which implies that the simple unweighted sample mean

$\bar{y}(s)$ is an unbiased estimator of \bar{Y} . We will refer to (1) as the expected proportional allocation (EPA) constraint.

The linear programming technique of Sitter and Skinner (1994) chooses a sampling design $p(s)$ which minimizes the expected lack of 'desirability' of the samples by solving the linear programming problem:

$$\min \sum_{s \in S_n} w(s) p(s) \quad (2)$$

subject to the constraint (1), where $w(s)$ is a loss function for the sample s , to be specified, and the $p(s)$ are the unknowns. Sitter and Skinner (1994) were exploiting the key observation of Rao and Nigam (1990, 1992) in the context of avoiding undesirable samples, that the objective function in (2) was linear in the $p(s)$'s (see also Lahiri and Mukerjee 2000).

In the objective function (2), the loss function $w(s)$ plays an important role. With a well defined $w(s)$, we have flexibility to explore the existence of an optimal solution to (2) within an economically sized S_n and, more importantly, to improve efficiency of estimation. Sitter and Skinner (1994) suggest choosing

$$w(s) = \sum_{i=1}^R (n_i(s) - n P_{i\cdot})^2 + \sum_{j=1}^C (n_j(s) - n P_{\cdot j})^2, \quad (3)$$

where $n_i(s) = \sum_j n_{ij}(s)$, $n_j(s) = \sum_i n_{ij}(s)$, $P_{i\cdot} = \sum_j P_{ij}$ and $P_{\cdot j} = \sum_i P_{ij}$. Obviously, the objective function (2) is actually $E(w(s))$ for any given design $p(s)$ and can be explained as the mean squared error of estimator \bar{y} under an analysis of variance model (see Sitter and Skinner 1994). Then by solving the above linear programming problem, one can obtain minimized MSE in the sense of ANOVA while maintaining the EPA property of the $n_{ij}(s)$. One should note that if a design with objective function equal to zero is obtained, then all margin constraints are met. This would typically only be the case with integer margins.

Sitter and Skinner (1994) suggest that one simple way to reduce the size of S_n is to restrict the actual values that n_{ij} can take to be either $\lfloor n P_{ij} \rfloor$ or $\lfloor n P_{ij} \rfloor + 1$, where $\lfloor n P_{ij} \rfloor$ is the greatest integer less than or equal to $n P_{ij}$. By denoting $\tilde{n}_{ij} = n_{ij} - \lfloor n P_{ij} \rfloor$ and $r_{ij} = n P_{ij} - \lfloor n P_{ij} \rfloor$, one can then impose

$$E(\tilde{n}_{ij}) = r_{ij}, \quad (4)$$

where $\tilde{n}_{ij} = 0$ or 1 and $0 \leq r_{ij} < 1$. Then the linear programming method can be applied to the \tilde{n}_{ij} and finally $\lfloor n P_{ij} \rfloor + \tilde{n}_{ij}$ can be used as the actual cell sample sizes. Therefore, without loss of generality, we will assume that

$$n_{ij} = 0, 1 \quad \text{and} \quad 0 \leq r_{ij} = n P_{ij} < 1. \quad (5)$$

2.2 Higher-way Stratification

The Sitter and Skinner (1994) approach extends straightforwardly to more stratifying factors by letting s denote the corresponding r -way array. The loss function would then include more terms, for example for three-way stratification equation (3) could be replaced by

$$w(s) = \gamma_1 \sum_{i=1}^{R_1} (n_{i..}(s) - nP_{i..})^2 + \gamma_2 \sum_{j=1}^{R_2} (n_{.j.}(s) - nP_{.j.})^2 + \gamma_3 \sum_{k=1}^{R_3} (n_{..k}(s) - nP_{..k})^2$$

in obvious notation, where γ_1, γ_2 and γ_3 might represent the relative importance of balancing on the three factors based on prior information (see Sitter and Skinner 1994).

2.3 Multi-stage Sampling

An important application of multi-way stratification is to the selection of primary sampling units (psu's) in multi-stage sampling, where it is more common to have several stratifying factors available.

In section 2.1, the inclusion probabilities of each unit are $E(n_{ij}(s)/N_{ij}) = n/N$. If psu's are selected with equal probability then the approach extends directly with the psu's the units and with the observed values of y replaced by unbiased estimators of the psu totals. However, if the psu's are to be selected with unequal probabilities, say nz_{ijk} for psu k in stratification cell $ij(z_{ijk}$ will typically equal $M_{ijk}/\sum_{ijk} M_{ijk}$, with M_{ijk} being some measure of size of psu k in cell ij), then the procedure can be easily modified by setting P_{ij} equal to $z_{ij}/z_{...}$, where $z_{ij} = \sum_k z_{ijk}$ and $z_{...} = \sum_{ijk} z_{ijk}$. Then, if $n_{ij}(s) > 0$, a sample of $n_{ij}(s)$ psu's in cell ij is selected by some probability proportional to z_{ijk} method.

2.4 An Example

The linear programming approach can be illustrated using the hypothetical example of Bryant *et al.* (1960) given in Table 1. First, this problem is simplified as shown in Table 2 to meet the assumption in (5). Then, a standard linear programming package is used to solve this reduced problem (2). Because integer margins of expected sample cell counts can be exactly matched by marginal totals of sample sizes $n_{i.}$ and $n_{.j.}$, which means that the loss function $w(s)$ can achieve a minimum value of zero, the objective function in (2) for this example is also minimized at zero. The optimal solution of this problem is given in Table 3. It should be noted that this solution has been converted back to match the original example shown in Table 1.

Table 2
Modified Example from Bryant *et al.* (1960)

Region	Type of Community			Total
	Urban	Rural	Metropolitan	
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

Table 3
Linear Programming Solution to Example
from Bryant *et al.* (1960)

s			$p(s)$			s			$p(s)$		
1	1	0	1	1	0	1	1	0			
1	0	0	0	0	1	0	0	1			
0	1	1	0.2	0	1	1	0.1	0	0	2	0.2
0	2	1		1	1	1		1	2	0	
1	0	1		1	1	0		1	1	0	
1	0	1		1	0	1		1	0	1	
0	1	0		0	1	0		0	0	1	
1	0	1	0.2	0	1	1	0.1	0	1	1	0.2
0	2	1		1	1	1		1	2	0	
1	1	0		1	1	0		1	1	0	

The linear programming method is simple and easy to use. Its main drawback is computational. The number of parameters in the resulting linear programming problem is the number of samples of size n from the $RC > n$ cells, $\binom{RC}{n}$, which becomes infeasibly large quite quickly. In the next section we will explore ways of improving the computational efficiency of the linear programming approach while maintaining all of its good properties.

3. THE LINEAR PROGRAMMING APPROACH
MADE PRACTICAL

The basic idea of the linear programming approach is to obtain an optimal sampling design in terms of the (minimum) expected lack of "desirability" of the sample by directly solving a linear programming problem with $p(s), s \in S_n$, as the unknowns while maintaining the EPA property. The only obstacle to this approach is that the number of elements in S_n is often very large and even with modern computing power it becomes difficult to carry out linear programming if the number of unknowns is large.

To reduce the magnitude of the computational task for this linear programming problem determined by the cardinality of S_n , we want to obtain a subset of S_n , say S_{n0} , which is nearly as representative as S_n but much smaller, and thus solve the following linear programming problem with a much smaller set of $p(s), s \in S_{n0}$, as the unknowns:

$$\min \sum_{s \in S_{n0}} w(s)p(s). \tag{6}$$

Hopefully, in this way we can easily deal with larger practical problems without losing the good properties of the linear programming approach.

3.1 Some Motivating Strategies

The above strategy is easy to state, but it turns out not to be entirely obvious how to go about it. In fact, there are several different directions we can explore to determine such a subset $S_{n0} \subset S_n$. In this section, we will describe a

basic method related to loss functions which was alluded to in Sitter and Skinner (1994) and describe how it modestly increases the size of problems that can be handled. We will then discuss some obvious directions to take which did not improve things much. By describing these misguided attempts, we motivate the eventual proposal.

The major flexibility of the linear programming approach is derived from the choice of loss function $w(s)$. Thus, it is natural for us to consider the loss function first when we try to improve the computational efficiency of this approach. By observing the objective function of the linear programming problem (2), we suspect that the loss function $w(s)$ as coefficients of unknowns $p(s)$ will not be very large when the objective function has been minimized. In other words, all positive $p(s)$ in an optimal sampling design will only be assigned to samples having small lack of “desirability”. Based on this observation, we hypothesize that the following subset might be a good replacement for S_n ,

$$S_{n_0} = \{s \in S_n : w(s) = \sum_{i=1}^R (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^C (n_{\cdot j}(s) - nP_{\cdot j})^2 \leq w_0\}, \quad (7)$$

where w_0 is a pre-determined positive constant. In the case of integer margins, one could even let $w_0 = 0$ and restrict to samples where the margins are matched. For example, the solution in Table 3 assigned positive probability to only 6 samples and for each of these the objective function was zero.

Lu (2000) develops nested linear programming strategies for solving this problem. For moderately sized problems such as 8×5 arrays (*i.e.*, 40 cells) this approach does well. However, for larger problems the size of resulting candidate sets becomes large very quickly, even in the integer margin case. Thus for large problems the technique faces the same problem as before—a huge candidate set that results in the difficulty of solving a linear programming problem with too many unknowns.

In reality, even a candidate sample set S_{n_0} of the form in (7) is far larger than necessary for us to find an optimal solution. What we really need is a smaller but fairly representative subset, where by “small” we mean small enough to make it *possible* to solve the resulting linear programming problem and by “representative” we mean containing elements which promise that this linear programming problem is *feasible*.

Before going on to describe our eventual proposed solution to this problem, we would like to introduce some naive methods of obtaining such a “representative subset” that turned out not to work well. These are not that useful in practice, but they did inspire our thinking in proposing a more sophisticated approach.

1) Two Stage Optimization: First of all, we could try to break S_{n_0} in (7) into many subsets which are small enough to be handled by linear programming respectively. Hopefully, optimal solutions from each of these smaller sets in the first stage optimization procedure can be combined to form the desired representative set of samples. Then we can just collect these optimal solutions together and apply linear programming once more. We applied this method to some simulated examples of size 6×6 , 7×7 , 8×8 and 9×9 as a method of preliminary investigation of its potential. Generally, in the first two cases the method worked very well and quickly, in the 8×8 case the method was time consuming and was not always able to obtain optimal solutions, and in the 9×9 case the method became infeasible.

2) Resampling from S_{n_0} : We could also randomly select a proportion, say 10%, of the S_{n_0} in (7) and hope this proportion is statistically representative of the complete set. Unfortunately, simulation results showed that the proportion obtained in this way is not “representative” enough, and the resulting linear programming problem often does not have any feasible solution. For example, the method of nested linear programming discussed previously was able to obtain matched integer margin solutions for simulated 8×5 arrays, however, these solutions were obtained much quicker by repeatedly sampling 10% of S_{n_0} and applying the Sitter and Skinner (1994) method to this set until a feasible solution was obtained. However, when slightly larger cases were considered the method took an inordinate amount of time before finding a feasible solution, and quickly became impractical.

There are two problems with both these approaches. First, the size of S_{n_0} becomes huge combinatorically and even complete enumeration becomes difficult. Having to first obtain S_{n_0} and then cutting the problem into pieces will either quickly outstrip the practical limits on linear programming due to the size of the pieces or create a huge number of pieces. Second, both of these strategies are not in any way attempting to avoid samples which are particularly bad choices for meeting the EPA constraints. The question is, is there any way we can generate a fairly “representative” candidate sample subset without choosing such “useless” samples or, more generally, can we select candidate samples in which the frequency of an entry’s appearance is more or less related to its desired expected sample counts?, and also can we do so without first having to enumerate a large S_{n_0} ? The general idea revolves around the fact that if we could randomly select a candidate subset directly from S_n without complete enumeration using an unequal probability selection procedure which simultaneously ensures that the objective function is minimized for every sample while ensuring that the EPA property is satisfied we will have solved the problem without resorting to linear programming at all. We have been working on finding such a selection procedure, but have yet to succeed. What we have been able to do is to develop such a proce-

ture with approximate EPA (AEPA). We can then use it to randomly generate a candidate subset of samples, S_{n_0} , and then apply a linear programming technique to this subset.

3.2 A Sampling Procedure with AEPA Property

In this section we first describe the approach as it applies to the case of integer margins. That is, the column totals, $n_{.j} = \sum_{i=1}^R r_{ij}$, and the row totals, $n_{i.} = \sum_{j=1}^C r_{ij}$, are integer valued. We go on to discuss how it can easily be adapted to the general case. In the linear programming approach, the goal is to minimize the expected lack of 'desirability' of the samples while maintaining the EPA property. We propose to accomplish this in two stages. First, we will develop an unequal probability selection procedure which selects samples which exactly match the integer margins and also have the AEPA property. We will then randomly generate a moderately sized set of such arrays and then apply a modified linear programming technique to this subset of all possible arrays. This will be repeated with larger and larger such sets. We will describe the sampling procedure and then we will discuss the modified linear programming technique.

Here is the basic idea for constructing such a sampling procedure: for a two-way table (assuming the expected cell sample sizes have been adjusted to lie between 0 and 1 as was done in going from Table 1 to 2), first we draw a sequence of population cells to produce $a_{11}, a_{12}, \dots, a_{1C}$ in the first row using an unequal probability without replacement sampling procedure based on the expected counts of that row, where $a_{ij} = 1$ if the ij -th cell is selected and = 0 otherwise. Then we draw $a_{i1}, a_{i2}, \dots, a_{iC}$ subsequently for $i > 1$ while keeping all $\sum_{k=1}^C a_{kj}$ less than or equal to the corresponding marginal column totals $n_{.j}$. The details of this sampling procedure are as follows:

Step 1: Randomly permute the rows and let $i = 1$. Given the first row of inclusion probabilities $r_{11}, r_{12}, \dots, r_{1C}$, draw a sample of $n_{1.}$ cells out of C in the first row stratum using an unequal probability without replacement sampling procedure; record the first row of samples in terms of indicator variables $a_{11}, a_{12}, \dots, a_{1C}$ as defined previously; let $A_j = a_{1j}$ for $j = 1, \dots, C$.

Step 2: Let $i = i + 1$

Step 2.1: For $j = 1, \dots, C$, do the following

- Let $R_j = \sum_{k=1}^i r_{kj}$
- If $R_j - A_j \leq 0$ let $a_{ij} = 0$,
- If $R_j - A_j \geq 1$ let $a_{ij} = 1$,

Step 2.2: Let $J = \{j : 0 < R_j - A_j < 1\}$ and $rtot = \sum_{j=1}^C r_{ij} - \#\{j : a_{ij} = 1\}$. If $rtot' > 0$ then $r_{ij}' = r_{ij} \times rtot / \sum_{j \in J} r_{ij}$, for $j \in J$. If there exists a $j_0 \in J$ such that $r_{ij_0}' > 1$ then let $a_{ij_0} = 1$ and go to Step 2.1. Otherwise go to Step 3.

Step 3: Draw a sample of $rtot$ cells from J using an unequal probability without replacement sampling procedure and r_{ij}' to get a_{ij} for $j \in J$.

Let $A_j = \sum_{k=1}^i a_{kj}$ for $j = 1, \dots, C$.

Step 4: If $i = R$, then stop; otherwise go to Step 2.

One aspect of this sampling procedure that should be noticed is that in Step 2, the way of re-calculating the i -th row of inclusion probabilities is not unique. However, the general rules that should be followed for this re-calculation are:

- $0 \leq r_{ij}' \leq 1$ and if $A_j = n_{.j}$, which means that there are enough units being selected from the j -th column, r_{ij}' should be set to 0; if $A_j = n_{.j} - (R - i + 1)$, which means that there will not be enough units to be selected for this column unless all of the remaining units are selected, r_{ij}' should be set to 1;
- keep $\sum_{j=1}^C r_{ij}' = \sum_{j=1}^C r_{ij} = n_{i.}$

The method extends easily to non-integer margins. We delay detailed discussion, however, to the sequel.

We can now use the above method to generate a candidate set, S_{n_0} , and apply the linear programming technique to this set. To see why we choose to modify the linear programming technique, realize that for the integer margin case every $s \in S_{n_0}$ already attains the minimum in (2) so that a direct application of linear programming amounts to determining whether there is a feasible solution or not. Thus, if we generate say an S_{n_0} of size 500 then 1,000 *etc.*, and the linear programming package continues to find no feasible solution we really do not know if we are getting closer to a solution or not. Instead we choose to turn the optimization around and solve a dual problem

$$\min_{p(s)} \sum_{i,j} \left| \sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} \right|. \quad (8)$$

We know that $w(s) = 0$ for all $s \in S_{n_0}$ and we are looking for a solution which yields a minimum of zero in (8). We have essentially switched the roles of the objective function and the EPA constraints in the original problem. The difficulty is that it is more difficult to use linear programming to handle (8). This can be done as follows. Set up constraints

$$\sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} + d_{ij} - e_{ij} = 0 \quad \text{for } i = 1, \dots, R \quad \text{and } j = 1, \dots, C, \quad (9)$$

where

$$d_{ij} \geq 0, e_{ij} \geq 0, d_{ij}e_{ij} = 0. \quad (10)$$

Then note that

$$\begin{aligned} \left| \sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} \right| &= \begin{cases} d_{ij} & \text{if } \sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} < 0 \\ e_{ij} & \text{if } \sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} \geq 0 \end{cases} \\ &= d_{ij} + e_{ij}. \end{aligned} \quad (11)$$

Thus, we can replace (8) by

$$\min_{p(s), d_{ij}, e_{ij}} \sum_{i,j} (d_{ij} + e_{ij}), \quad (12)$$

subject to

$$\sum_{s \in S_{n_0}} n_{ij}(s)p(s) - r_{ij} + d_{ij} - e_{ij} = 0, d_{ij}, e_{ij}, p(s) \geq 0, d_{ij}, e_{ij} = 0. \quad (13)$$

3.3 Some Illustrating Examples with Integer Margins

In this section, two examples will be used to illustrate the sampling procedure. The first with a 10×8 array is described in detail to show the whole procedure. The second with a larger size (20×15) is given to demonstrate the size of problem that this method can handle (this is near the limit of the problem the proposed method can realistically handle). Any unequal probability without replacement sampling procedure can be used within the method. In Example 1 below, we chose to use the random grouping method of Rao, Hartley and Cochran (1962), since it is simple and we really only need to approximately match the selection probabilities, which it does. However, the Rao-Hartley-Cochran method only works well up to problems of moderate size. In Examples 2 and 3 one should use a method which exactly matches the selection probabilities. There are many such available, but we chose to use one developed in Lu (2000).

Example 1. 10×8 array with integer margins: A two-way stratification problem with expected sample cell counts and sample size is given in Table 4.

Table 4
Expected Sample Cell Counts Under Proportionate Stratification ($n = 40$)

Row No.	Column No.								Marginal Row Total
	1	2	3	4	5	6	7	8	
1	0.41	0.55	0.58	0.80	0.23	0.61	0.70	0.12	4
2	0.52	0.15	0.07	0.90	0.28	0.10	0.37	0.61	3
3	0.72	0.15	0.65	0.73	0.39	0.34	0.85	0.17	4
4	0.70	0.55	0.46	0.10	0.41	0.05	0.24	0.49	3
5	0.07	0.63	0.45	0.81	0.52	0.02	0.70	0.80	4
6	0.61	0.33	0.79	0.21	0.02	0.61	0.67	0.76	4
7	0.88	0.48	0.73	0.69	0.44	0.64	0.86	0.28	5
8	0.22	0.14	0.85	0.37	0.69	0.45	0.49	0.79	4
9	0.85	0.44	0.80	0.76	0.31	0.71	0.60	0.53	5
10	0.02	0.58	0.62	0.63	0.71	0.47	0.52	0.45	4
Marginal Col Total	5	4	6	6	4	4	6	5	40

The basic steps of our sampling design are illustrated as follows:

Step 1. Obtain a representative candidate sample subset S_{n_0} by using proposed sampling procedure with AEPA property to draw, say 500, samples (obtained within 3 minutes). The sample proportion of each cell is shown in Table 5, which can be compared to Table 4 to see how close these are to satisfying the EPA property.

Step 2. Solve the linear programming problem given by (12) and (13) to obtain

$$\min_{p(s), s \in S_{n_0}} \sum_{i,j} \left| \sum_s n_{ij}(s)p(s) - nP_{ij} \right|. \quad (14)$$

If the objective value of (14) is greater than zero, repeat Step 1 with a larger set S_{n_0} . If the objective value of (14) is zero, stop, an optimal solution has been obtained.

Table 5
Sample Cell Counts Under Prop. Stratification ($n = 40$)

Row No.	Column No.								Marginal Row Total
	1	2	3	4	5	6	7	8	
1	0.408	0.554	0.582	0.776	0.250	0.594	0.734	0.102	4
2	0.554	0.150	0.062	0.916	0.280	0.122	0.366	0.550	3
3	0.690	0.144	0.638	0.720	0.402	0.360	0.838	0.208	4
4	0.692	0.542	0.452	0.120	0.416	0.044	0.260	0.474	3
5	0.060	0.602	0.446	0.814	0.568	0.016	0.708	0.786	4
6	0.558	0.348	0.780	0.216	0.012	0.634	0.682	0.770	4
7	0.866	0.480	0.734	0.676	0.470	0.664	0.842	0.268	5
8	0.254	0.158	0.848	0.400	0.654	0.412	0.490	0.784	4
9	0.870	0.418	0.830	0.772	0.292	0.692	0.624	0.502	5
10	0.026	0.564	0.636	0.658	0.714	0.416	0.500	0.486	4
Marginal Col Total	5	4	6	6	4	4	6	5	40

In this example, a candidate subset S_{n_0} with 500 samples was sufficient to get objective value of 0.

Example 2. 20×15 array with integer margins: In this example, a 20×15 array with integer margins is given in Table 6.

The actual computation steps are given as follows:

First Iteration:

Step 1. Draw 500 samples to form S_{n_0} .

Step 2. The objective value of (14) is 0.1659.

Second Iteration:

Step 1. Draw 500 samples to add to S_{n_0} .

Step 2. The objective value of (14) is 0. The final sampling design is attained.

This procedure took approximately 30-60 seconds using a Fortran program on a Sun Ultra 10 workstation.

3.4 Extension to Non-Integer Margins

The method extends easily to non-integer margins. Merely replace n_{ij} throughout the algorithm by n_{ij}^* , which takes value $\lfloor r_{ij} \rfloor + 1$ with probability $\alpha = r_{ij} - \lfloor r_{ij} \rfloor$ and takes value $\lfloor r_{ij} \rfloor$ with probability $1 - \alpha$. The only additional difficulty is that $E[w(s)]$ cannot attain zero. Thus, we do not have an obvious lower-bound reference point to ascertain whether we are close to the best solution or not. However, the above randomization strategy ensures that for every obtained AEPA sample we have

$$|n_{i.}(s) - r_{i.}| < 1 \quad \text{and} \quad |n_{.j}(s) - r_{.j}| < 1$$
$$\text{for } i = 1, \dots, R, j = 1, \dots, C. \quad (15)$$

This together with the EPA property, $E[n_{ij}(s)] = \sum_s n_{ij}(s)p(s) = r_{ij}$ implies that the lack of desirability function $w(s)$ defined in (3) has a constant expectation

$$E[w(s)] = \sum_i (r_{i.} - \lfloor r_{i.} \rfloor)(1 + \lfloor r_{i.} \rfloor - r_{i.})$$
$$+ \sum_j (r_{.j} - \lfloor r_{.j} \rfloor)(1 + \lfloor r_{.j} \rfloor - r_{.j}). \quad (16)$$

The proof of this is given in Appendix 1. Thus, if (14) attains zero under the above strategy then the resulting solution will yield minimum $E[w(s)]$ as in (16).

Example 3. 27 × 3 real example with non-integer margins: We will illustrate the method using a real example from environmental health (Vihma 1981). This study was concerned with occupational health of workers in various industries in Finland. The population chosen for study consisted of 1,430 small industrial workplaces (5-49 employees) totalling 22,893 employees in Uusimaa, the southern most and most industrialized province of Finland. The primary sampling units were the workplaces and a sample of $n=100$ such were desired. This was all that could be afforded given the cost of the eventual survey. The

workplaces were stratified by two stratification variables: type of industry (27 categories) and number of employees (3 categories). The expected sample cell counts under proportionate stratification are given in Table 7. The actual sampling scheme used in this study was based on the method of Bryant *et al.* (1960) after some grouping strata as it was the only method available at the time of this study.

We applied our method to this problem. The minimum achievable $E[w(s)]$ using our proposed strategy is 5.0418. The actual computation steps were as follows:

First Iteration:

- Step 1.** Draw 500 samples to form S_{n_0} , randomly generating the n_{ij}^* independently for each sample.

Step 2. The objective value of (14) is 0.45088.

Second Iteration:

- Step 1.** Draw 500 samples to add to S_{n_0} .

Step 2. The objective value of (14) is 0. The final sampling design is attained and achieved the minimum value $E[w(s)] = 5.0418$.

This procedure took approximately 30 seconds using a Fortran program on a Sun Ultra 10 workstation.

Table 6
Expected Sample Cell Counts Under Proportionate Stratification ($n=151$)

0.73	0.58	0.08	0.59	0.69	0.84	0.04	0.17	0.27	0.80	0.02	0.84	0.79	0.03	0.53	7
0.43	0.39	0.35	0.57	0.35	0.38	0.47	0.53	0.39	0.96	0.52	0.27	0.68	0.40	0.31	7
0.73	0.25	0.15	0.73	0.48	0.32	0.91	0.49	0.03	0.61	0.14	0.61	0.73	0.25	0.87	7
0.13	0.28	0.35	0.60	0.26	0.38	0.37	0.39	0.71	0.01	0.93	0.72	0.30	0.66	0.91	7
0.32	0.06	0.86	0.47	0.80	0.93	0.96	0.30	0.65	0.72	0.67	0.54	0.51	0.77	0.44	9
0.12	0.78	0.81	0.34	0.28	0.02	0.89	0.41	0.94	0.82	0.37	0.81	0.85	0.51	0.05	8
0.48	0.51	0.50	0.62	0.35	0.11	0.85	0.78	0.29	0.39	0.69	0.07	0.67	0.78	0.91	8
0.86	0.41	0.11	0.17	0.75	0.89	0.48	0.48	0.91	0.20	0.53	0.67	0.34	0.19	0.01	7
0.81	0.00	0.13	0.93	0.36	0.12	0.19	0.86	0.33	0.04	0.79	0.69	0.56	0.37	0.82	7
0.82	0.22	0.54	0.82	0.61	0.46	0.74	0.33	0.24	0.53	0.41	0.18	0.30	0.03	0.77	7
0.95	0.60	0.35	0.33	0.95	0.43	0.06	0.63	0.71	0.02	0.55	0.23	0.87	0.21	0.11	7
0.96	0.65	0.96	0.83	0.41	0.58	0.49	0.27	0.74	0.88	0.93	0.46	0.60	0.13	0.11	9
0.83	0.54	0.05	0.96	0.79	0.70	0.33	0.81	0.86	0.45	0.45	0.84	0.29	0.30	0.80	9
0.75	0.65	0.63	0.04	0.32	0.36	0.38	0.80	0.50	0.23	0.37	0.23	0.85	0.69	0.20	7
0.79	0.31	0.55	0.26	0.04	0.05	0.91	0.11	0.43	0.79	0.14	0.64	0.44	0.48	0.06	6
0.23	0.92	0.81	0.42	0.49	0.10	0.74	0.56	0.24	0.47	0.34	0.57	0.60	0.56	0.95	8
0.13	0.77	0.65	0.66	0.05	0.23	0.58	0.74	0.19	0.94	0.26	0.75	0.16	0.71	0.18	7
0.31	0.01	0.60	0.38	0.01	0.55	0.70	0.72	0.20	0.87	0.55	0.82	0.77	0.44	0.07	7
0.63	0.67	0.21	0.02	0.16	0.68	0.14	0.17	0.95	0.78	0.58	0.55	0.94	0.96	0.56	8
0.99	0.40	0.31	0.26	0.85	0.87	0.77	0.75	0.42	0.49	0.76	0.51	0.75	0.53	0.34	9
12	9	9	10	9	9	11	10	10	11	10	11	12	9	9	151

Table 7

Occupational Health Survey, Vihma (1981) Expected Sample Cell Counts Under Proportionate Stratification ($n = 100$)

Type of Industry	Number of Personnel			
	5-9	10-19	20-49	r_i
Food products	2.38	3.56	3.78	9.72
Food	0.35	0.14	0.56	1.05
Beverage	0.14	0.07	0.21	0.42
Textiles	1.33	1.26	1.46	4.05
Apparel	3.15	3.71	2.09	8.95
Leather	0.56	0.14	0.07	0.77
Footwear	0.07	0.07	0.21	0.35
Wood Products	2.37	1.89	0.91	5.17
Furniture	1.33	0.84	0.91	3.08
Paper Products	0.42	0.49	0.42	1.33
Printing	7.20	6.01	4.20	17.41
Industrial Chemicals	0.56	0.35	0.28	1.19
Chemical Products	1.82	1.54	1.53	4.89
Petroleum	0.14	0.07	0.00	0.21
Misc Coal and Petrol.	0.07	0.07	0.14	0.28
Rubber Products	0.14	0.21	0.07	0.42
Plastic Products	1.40	1.05	1.19	3.64
Glass Products	0.42	0.21	0.21	0.84
Non-Metal Minerals	1.12	0.98	0.84	2.94
Iron & Steel	0.14	0.07	0.35	0.56
Nonferrous Metal	0.35	0.14	0.28	0.77
Fabricated Metal	4.96	4.06	2.59	11.61
Machinery	2.80	1.96	3.21	7.97
Electrical	1.89	1.60	1.33	4.82
Transport Equipment	0.84	0.84	0.84	2.52
Scientific Equipment	0.56	0.42	0.49	1.47
Manufacturing Industries	1.68	0.91	0.98	3.57
n_j	38.19	32.66	29.15	100.00

5. CONCLUDING REMARKS

We propose a method for two-way stratification which extends the applicability of the linear programming approach of Sitter and Skinner (1994) to much larger problems. The method focuses on how to construct a small "representative" candidate sample set by using an unequal probability sampling procedure which generates candidate samples which nearly meet the AEPA constraints of the linear programming problem and then applying the linear programming method to this much smaller set.

It should be noted that the linear programming method extends easily to stratified multi-stage designs. Since there is no fundamental difference between the original linear programming approach and the extension proposed here, this is still true of the proposed method. In the same spirit, one can view discussion on issues around variance estimation of the resulting estimators in Sitter and Skinner (1994) as well.

One should also note that once one restricts to bracketing integers around the nP_{ij} 's, the problem is related to a

controlled rounding problem (see Kelly, Golden and Assad 1993, and references therein), though we do not explore this aspect here.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

APPENDIX 1

Proof of (16): $n_i(s) - \lfloor r_i \rfloor \sim \text{Bernoulli}(r_i - \lfloor r_i \rfloor)$ and has variance $(r_i - \lfloor r_i \rfloor)(1 + \lfloor r_i \rfloor - r_i)$. This implies

$$\begin{aligned} \sum_s (n_i(s) - r_i)^2 p(s) &= E(n_i(s) - r_i)^2 V(n_i(s)) \\ &= V(n_i(s) - \lfloor r_i \rfloor) \\ &= (r_i - \lfloor r_i \rfloor)(1 + \lfloor r_i \rfloor - r_i), \end{aligned}$$

and by similar argument that $\sum_s (n_j(s) - r_j)^2 p(s) = (r_j - \lfloor r_j \rfloor)(1 + \lfloor r_j \rfloor - r_j)$.

Therefore, with $w(s)$ defined in (3),

$$\begin{aligned} E[w(s)] &= \sum_s w(s)p(s) = \sum_s \left\{ \sum_i (n_i(s) - r_i)^2 + \sum_j (n_j(s) - r_j)^2 \right\} p(s) \\ &= \sum_i \sum_s (n_i(s) - r_i)^2 p(s) + \sum_j \sum_s (n_j(s) - r_j)^2 p(s) \\ &= \sum_i (r_i - \lfloor r_i \rfloor)(1 + \lfloor r_i \rfloor - r_i) + \sum_j (r_j - \lfloor r_j \rfloor)(1 + \lfloor r_j \rfloor - r_j). \end{aligned}$$

REFERENCES

- BRYANT, E.C., HARTLEY, H.O. and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*. 55, 105-124.
- CAUSEY, B.D., COX, L.H. and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- ENGLE, M., MARSDEN, G. and POLLOCK, S.W. (1971). Child work and social class. *Psychiatry*. 34, 140-150.
- GOODMAN, R., and KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*. 45, 350-372.
- HESS, I., RIEDEL, D.C. and FITZPATRICK, T.B. (1976). *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, second edition.
- JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*. 65, 776-795.
- JESSEN, R.J. (1973). Some properties of probability lattice sampling. *Journal of the American Statistical Association*. 68, 20-28.
- JESSEN, R.J. (1975). Square and cubic lattice sampling. *Biometrics*. 31, 449-471.

- KELLY, J.K., GOLDEN, B.L. and ASSAD, A.A. (1993). The controlled rounding problem: complexity and computational experience. *European Journal of Operational Research*. 65, 207-217.
- LAHIRI, P., and MUKERJEE, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistical Sinica*. 10, 1171-1178.
- LU, W. (2000). Multi-way stratification by linear programming made practical. M.Sc. Thesis, Simon Fraser University.
- RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. Serie B*, 24, 482-491.
- RAO, J.N.K., and NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*. 77, 807-814.
- RAO, J.N.K., and NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*. 60, 89-98.
- SITTER, R.R., and SKINNER, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*. 20, 65-73.
- SKINNER, C.J., HOLMES, D.J. and HOLT, D. (1994). Multiple frame sampling for multiple stratification. *International Statistical Review*. 62, 333-347.
- VIHMA, T. (1981). Health hazards and stress factors in small industry-Prevalence study in the province of Uusimaa with special reference to the type of industry and the occupational title as classifications for the description of occupational health problems. *Scandinavian Journal of Work, Environment and Health*. 7, Suppl. 3, 1-149.
- WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*. 32, 150-164.

On the Use of Generalized Inverse Matrices in Sampling Theory

ROBERT H. RENSEN and GERARD H. MARTINUS¹

ABSTRACT

In theory, it is customary to define general regression estimators in terms of full-rank weighting models, *i.e.*, the design matrix that corresponds to the weighting model is of full rank. For such weighting models, it is well known that the general regression weights reproduce the (known) population totals of the auxiliary variables involved. In practice, however, the weighting model often is not of full rank, especially when the weighting model is for incomplete post-stratification. By means of the theory of generalized inverse matrices, it is shown under which circumstances this consistency property remains valid. As a non-trivial example we discuss the consistent weighting between persons and households as proposed by Lemaître and Dufour (1987). We then show how the theory is implemented in Bascula.

KEY WORDS: Bascula; General regression estimator; Weighting.

1. INTRODUCTION

Weighting methods that are based on the general regression estimator are commonly used in sample surveys to adjust for both sampling error and non-sampling error, see *e.g.* Bethlehem and Keller (1987) and Särndal, Swensson, and Wretman (1992). One complication in the use of general regression estimators, however, is that many weighting models are based on incomplete post-stratification, resulting in design matrices that are not of full rank. Usually, this problem is solved by using a reduced design matrix. Such a reduced design matrix can be constructed by deleting redundant columns and properly adjusting the population totals. Often, the redundancy can be recognized rather easily beforehand by the specification of the weighting model. However, for some weighting models such a redundancy check may be impractical.

For example, suppose that we have a post-stratification based on the complete crossing between two categorical variables A and B , with known counts for the population of each cell. We may obtain small sample counts or no sample in some cells. Then we may derive new classifications, A' from A and B' from B , by merging categories, and define the following more parsimonious scheme: $A + B + A' \times B'$. According to this incomplete post-stratification we simultaneously calibrate on three sets of counts, namely the marginal counts of A , the marginal counts of B , and the cell counts of $A' \times B'$. Since A and A' (and also B and B') appear in different weighting terms, it is difficult to recognize redundancy by the specification of the weighting model. This paper gives the theoretical background, which is based on generalized inverse matrices, of reducing such a design matrix.

In section 2 we briefly describe some properties of generalized inverse matrices. In section 3 we define the general regression estimator for weighting models that need not be of full rank. Given a regularity condition that can be

nicely interpreted in a calibration estimation context (see Deville and Särndal 1992) it is shown that this estimator is invariant with respect to the choice of the generalized inverse. At the end of section 3 the fulfillment of this regularity condition is discussed for some well-known weighting models, such as incomplete post-stratification and consistent weighting between persons and households. In section 4 we describe the algorithm, which is implemented in Bascula (see Nieuwenbroek 1997; Rensen, Nieuwenbroek and Slootbeek 1997) for calculating the regression weights. Finally, in section 5 we briefly discuss the weighting model of the Dutch Labour Force Survey.

2. GENERALIZED INVERSE MATRICES

We are mainly interested in the use of generalized inverses within the framework of the general regression estimator. Hence, we only give some properties of a generalized inverse of the form $X' \Lambda X$, where Λ is a diagonal matrix of order $n \times n$ with strictly positive diagonal entries and X a design matrix of order $n \times p$ that results from the weighting model. For a more extensive discussion on generalized inverse matrices we refer to Searle (1971) and Rao (1973).

Before giving these properties, we briefly review the definition of a generalized inverse. Consider a $p \times q$ matrix A of any rank and let $Ax = y$ be a system of consistent equations, *i.e.*, any linear relationship existing among the rows of A also exists among the corresponding elements of y . A generalized inverse of A is a $q \times p$ matrix A^- such that $x = A^-y$ is a solution of this system of equations. It is easy to verify that the existence of A^- implies $AA^-A = A$ (choose y as the i -th column of A). Conversely, if A^- satisfies $AA^-A = A$ and $Ax = y$ is consistent, then $A(A^-y) = A(A^-Ax) = Ax = y$ and hence A^-y is a solution. Thus, as an alternative definition, a generalized

¹ Robert H. Rensen and Gerard H. Martinus, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands.

inverse matrix of A is any matrix A^- such that $AA^-A = A$.

Now, if G denotes a generalized inverse of $X' \Lambda X$, then the following properties of G are proven in Searle (1971) for $\Lambda = I_n$:

- (P1) G' is also a generalized inverse of $X' \Lambda X$,
- (P2) $XXG' \Lambda X = X$ i.e., $G' \Lambda$ is a generalized inverse of X ,
- (P3) XXG' is invariant to the choice of G ,
- (P4) $XXG' = XG'X'$ whether G is symmetric or not.

The proofs of (P1) to (P4) for diagonal Λ are almost identical to those of Searle (1971, chapter 1.5, theorem 7) and therefore not repeated here.

3. THE GENERAL REGRESSION ESTIMATOR

Consider a finite population U of N units from which a sample S of n units is drawn without replacement. Let π_k denote the first order inclusion probability of the k -th unit, $k = 1, \dots, N$. We associate with each unit a vector of study variables y_k . Then, the data matrix for the sampled units is given by $Y_S = (y_1, \dots, y_n)'$. We distinguish between study variables with known population totals (auxiliary variables) and study variables with unknown population totals. The start in the definition of a general regression estimator (Särndal *et al.* 1992) is the specification of the weighting model, i.e., the choice of the set of auxiliary variables to be used in the estimation. Denoting this specific set of p variables by x , we call the $n \times p$ matrix $X_S = (x_1, \dots, x_n)'$ the design matrix, which is, by definition, a column subset of Y_S . The vector of known population totals of x is denoted by t_x . Let $x_{HT} = \sum_{k \in S} \pi_k^{-1} x_k$ denote the Horvitz-Thompson estimator for t_x , then, given x , the general regression estimator of the vector of population totals of the i -th study variable $y_k^{(i)}$ is defined as

$$\hat{t}_{\text{greg}}^{(i)} = y_{HT}^{(i)} + \hat{B}'(t_x - x_{HT}) \quad (1)$$

with

$$\hat{B} = G_S X_S' \Lambda_S Y_S^{(i)}.$$

In terms of regression weights, this general regression estimator can also be written as

$$\hat{t}_{\text{greg}}^{(i)} = \sum_{k \in S} w_k y_k^{(i)} \quad (2)$$

with

$$w_k = \pi_k^{-1} + \lambda_k x_k' G_S (t_x - x_{HT}).$$

Here, G_S denotes a generalized inverse of $X_S' \Lambda_S X_S$ and $\Lambda_S = \text{diag}(\lambda_1, \dots, \lambda_n)$ is some diagonal matrix with strictly positive entries.

Like the weighting model, the diagonal matrix Λ_S has to

be specified by the user. Often, one takes $\Lambda_S = \Pi_S^{-1} \Sigma_S^{-1}$, where $\Pi_S = \text{diag}(\pi_1, \dots, \pi_n)$ and $\Sigma_S = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ with σ_k^2 interpreted as the variance of independent random variables of which some of the study variables are supposed to be the outcome according to some super-population model, see Särndal *et al.* (1992). It is required that all σ_k^2 be known up to a common scale factor. An important special case is $\sigma_k^2 = \sigma^2$, i.e., all the modeled variances are the same. This results in the regression estimator proposed by Bethlehem and Keller (1987). If the population units represent households (of size m_k) and if we take $\sigma_k^2 = m_k \sigma^2$ we arrive at the estimator proposed by Lemaître and Dufour (1987) to obtain consistent weights between person and households. From a different point of view, Alexander (1987) derived the GLS-P estimate, which results in essentially the same estimator.

Below we show that the regression weights are invariant to the choice of G_S . To that purpose we make the following assumption:

- (A1) there exists a n -vector w such that $X_S' w = t_x$.

Clearly, this assumption states that $X_S' w = t_x$ is a system of consistent equations. It is interesting to note that this system precisely corresponds to the set of calibrations equations when considering the general regression estimator as a special case of the calibration estimator (see e.g. Deville and Särndal 1992). If $X_S' w = t_x$ is a system of consistent equations, then so is $X_S' v = (t_x - x_{HT})$. This is easily seen by taking $v = w - d_S$ with $d_S = (\pi_1^{-1}, \dots, \pi_n^{-1})'$. The invariance of the regression weights to the choice of G_S , and hence the invariance of the general regression estimator can be shown as follows. Let F_S be some other generalized inverse of $X_S' \Lambda_S X_S$, different from G_S . Then, we have

$$\begin{aligned} X_S' G_S (t_x - x_{HT}) &= X_S' G_S X_S' v && \text{by (A1)} \\ &= X_S' F_S X_S' v && \text{by (P3)} \\ &= X_S' F_S (t_x - x_{HT}). && \text{by (A1)} \end{aligned}$$

So, it holds that $x_k' G_S (t_x - x_{HT})$ is invariant to G_S for all $k \in S$, implying that the regression weights are invariant to the choice G_S .

The fact that these weights reproduce the population totals of the auxiliary variables follows from the following series of equations:

$$\begin{aligned} \sum_{k \in S} w_k x_k &= x_{HT} + \sum_{k \in S} x_k \lambda_k x_k' G_S (t_x - x_{HT}) \\ &= x_{HT} + (X_S' \Lambda_S X_S) G_S (t_x - x_{HT}) \\ &= x_{HT} + (X_S' \Lambda_S X_S) G_S X_S' v && \text{by (A1)} \\ &= x_{HT} + X_S' v && \text{by (P2) and (P4)} \\ &= x_{HT} + (t_x - x_{HT}) = t_x. && \text{By (A1)} \end{aligned}$$

We close this section by having a closer look at the stated assumption for some well-known weighting models. In case of post-stratification in which the weighting model is described by a complete crossing of categorical variables, (A1) has a simple interpretation. Namely (A1) is satisfied if and only if empty post-strata in the sample correspond to empty post-strata in the population. Next, we consider incomplete post-stratification in which the weighting model consists of several terms, each term describing a complete crossing of categorical variables and so each term corresponding to a post-stratification. Then, a necessary condition for (A1) to be satisfied is that empty post-strata in the sample correspond to empty post-strata in the population for each of these terms. Unfortunately, this condition is not sufficient. For example, inconsistencies may still occur when we attempt to calibrate on a number of complete crossings larger than the sample size.

The assumption is less straightforward in case of consistent weighting between persons and households (see e.g. Lemaître and Dufour 1987). This is due to the redefinition of the auxiliary variable. For example, if \mathbf{x}_k is a variable defined at the person level, and from this variable a new variable is defined on the household level, say \mathbf{z}_k , then (A1) should be defined in terms of $\mathbf{Z}_S = (\mathbf{z}_1, \dots, \mathbf{z}_n)^t$ instead of \mathbf{X}_S , i.e., (A1) is satisfied if there exists an n -vector \mathbf{w} such that $\mathbf{Z}_S' \mathbf{w} = \mathbf{t}_x$. In many (regular) situations, the linear manifold spanned by \mathbf{Z}_S will coincide with the linear manifold spanned by \mathbf{X}_S . In such situations the method of Lemaître and Dufour does not affect the validity of (A1). However, in specific cases this may not be true. The following simplified example illustrates this.

Let \mathbf{x}_k denote sex of the k -th person, say $\mathbf{x}_k = (0, 1)^t$ if the k -th person is a female and $\mathbf{x}_k = (1, 0)^t$ if the k -th is a male. According to the method of Lemaître and Dufour (1987), let \mathbf{z}_k denote the j -th household mean for \mathbf{x}_k whenever k belongs to the j -th household. Furthermore, let the population consists of N_1 males and N_2 females, from which a sample of 10 households is drawn. Suppose that each sampled household consists of two persons, namely one male and one female. This gives $\mathbf{z}_k = (1/2, 1/2)^t$ for all $k \in S$. For this example the linear manifold spanned by \mathbf{Z}_S is a linear subspace of the linear manifold spanned by \mathbf{X}_S . If $N_1 = N_2$ then (A1) is satisfied. Otherwise, if $N_1 \neq N_2$ then (A1) is not satisfied. Especially, when the method of Lemaître and Dufour is applied on a relatively large weighting model, the linear manifold spanned by \mathbf{Z}_S may be a proper subspace of the linear manifold spanned by \mathbf{X}_S . Then, (A1) only is satisfied if \mathbf{t}_x accidentally belongs to this subspace.

4. CALCULATING THE REGRESSION WEIGHTS IN BASCULA

In the previous section we have shown that the general regression weights $w_k = \pi_k^{-1} + \lambda_k \mathbf{x}_k' \mathbf{G}_S (\mathbf{t}_x - \mathbf{x}_{HT})$ are

invariant to the choice of \mathbf{G}_S . In this section we show how to compute these weights. To do so, we start with the Cholesky decomposition of the positive (semi) definite matrix $\mathbf{X}_S' \Lambda_S \mathbf{X}_S$, see Seber (1977, page 322). If \mathbf{X}_S is of full rank, then $\mathbf{X}_S' \Lambda_S \mathbf{X}_S$ is positive definite and it can be expressed uniquely in the form $\mathbf{X}_S' \Lambda_S \mathbf{X}_S = \mathbf{U}' \mathbf{U}$, where \mathbf{U} is an upper triangular matrix with positive diagonal elements. Let a_{ij} denote the ij -th element of $\mathbf{X}_S' \Lambda_S \mathbf{X}_S$, then \mathbf{U} can be computed, row by row, according to

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2} \quad \text{for } i = 1, \dots, p$$

and

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj}}{u_{ii}} \quad \text{for } j = i+1, \dots, p. \quad (3)$$

If \mathbf{X}_S has rank $r < p$, then an application of (3) will give r non-zero and $p-r$ zero diagonal elements of \mathbf{U} . If we find a zero diagonal element then we put its corresponding row and column elements at zero. Subsequently, by elementary row and column interchanges, we obtain the following upper triangular matrix:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Accordingly to the elementary row and column interchanges we also interchange the elements of \mathbf{X}_S and $(\mathbf{t}_x - \mathbf{x}_{HT})$: $\mathbf{X}_S \mathbf{E}' = (\mathbf{X}_{1S} \mathbf{X}_{2S})$ and

$$\mathbf{E}(\mathbf{t}_x - \mathbf{x}_{HT}) = \begin{pmatrix} (\mathbf{t}_{1x} - \mathbf{x}_{1HT}) \\ (\mathbf{t}_{2x} - \mathbf{x}_{2HT}) \end{pmatrix},$$

where, by construction, \mathbf{X}_{1S} is of full rank and \mathbf{E} is a non-singular matrix of order $p \times p$. But, since

$$\mathbf{G}_S' = \begin{pmatrix} (\mathbf{X}_{1S}' \Lambda_S \mathbf{X}_{1S})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^{-1} (\mathbf{U}_1')^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

is a generalized inverse of $(\mathbf{X}_{1S} \mathbf{X}_{2S})' \Lambda_S (\mathbf{X}_{1S} \mathbf{X}_{2S})$, we have that $\mathbf{G}_S = \mathbf{E}' \mathbf{G}_S' \mathbf{E}$ is a generalized inverse of $\mathbf{X}_S' \Lambda_S \mathbf{X}_S$. Inserting this generalized inverse into $w_k = \pi_k^{-1} + \lambda_k \mathbf{x}_k' \mathbf{G}_S (\mathbf{t}_x - \mathbf{x}_{HT})$ gives

$$w_k = \pi_k^{-1} + \lambda_k (\mathbf{x}_{1k}' \mathbf{x}_{2k}') \mathbf{G}_S' \begin{pmatrix} (\mathbf{t}_{1x} - \mathbf{x}_{1HT}) \\ (\mathbf{t}_{2x} - \mathbf{x}_{2HT}) \end{pmatrix} \\ = \pi_k^{-1} + \lambda_k \mathbf{x}_{1k}' \mathbf{U}_1^{-1} (\mathbf{U}_1')^{-1} (\mathbf{t}_{1x} - \mathbf{x}_{1HT}),$$

which is computed as follows. First $\mathbf{z} = (\mathbf{U}_1')^{-1} (\mathbf{t}_{1x} - \mathbf{x}_{1HT})$ is computed by solving the lower triangular system $\mathbf{U}_1' \mathbf{z} = (\mathbf{t}_{1x} - \mathbf{x}_{1HT})$. Thereafter $\mathbf{u} = \mathbf{U}_1^{-1} \mathbf{z}$ is computed by solving the upper triangular system $\mathbf{U}_1 \mathbf{u} = \mathbf{z}$. Once

$\mathbf{u} = \mathbf{U}_1^{-1}(\mathbf{U}_1')^{-1}(\mathbf{t}_{1x} - \mathbf{x}_{\text{IHT}})$ is computed it is a simple matter to compute w_k .

5. THE DUTCH LABOUR FORCE SURVEY

To illustrate some of the issues stated in this paper, we briefly discuss the weighting model of the Dutch Labour Force Survey (LFS) of 1987 up to 2000. The target population of this survey consisted of the non-institutional population residing in the Netherlands and its sampling design was based on a stratified three-stage sampling with households as ultimate sampling units. For details we refer to Nieuwenbroek and Van der Valk (1996). Five categorical variables were involved into the weighting model, namely Sex (2 categories), Age (12 categories), Marital Status (2 categories), Region (15 categories), and Nationality (2 categories). Mainly based on consistency requirements, the desired weighting model was

Sex \times Age \times Marital Status \times Region \times Nationality.

However, this weighting model resulted in too many small cell counts, which gave unstable estimators. Therefore, the reduced model

(Sex \times Age \times Marital Status \times Region)
+ (Sex \times Age* \times Region \times Nationality)

was used instead, where Age* (2 categories) was obtained by grouping the categories in Age. This reduced weighting model resulted in a design matrix not of full rank for two reasons, namely 1) some columns of the design matrix completely consisted of zeros due to impossible combinations of the categorical variables and 2) there were linear combinations between the columns of the design matrix.

Now, the first kind of redundancy can be easily traced. If such columns are found, then their corresponding population totals should be zero. Bascula carries out a check on this condition. The second kind of redundancy is more difficult to trace. Linear combinations between columns may arise because one variable is incorporated into several weighting terms. For example, sex and region appear in both weighting terms of the LFS weighting model. The resulting linear combinations can be recognized beforehand by the name of the variable. For the age-variable, which also appears in both weighting terms, such a redundancy check beforehand is less obvious. These latter kinds of redundancy are traced by means of the Cholesky decomposition. Naturally, if any linear combinations are found, either by name beforehand or by the Cholesky decomposition, then the same linear combinations should also exist between the vector of population totals. Bascula also checks this condition.

ACKNOWLEDGEMENT

The authors wish to thank Jan van den Brakel, Nico Nieuwenbroek, and Jeroen Pannekoek for their careful reading and helpful comments. The authors also wish to thank the referee for his careful reading and valuable suggestions. Especially his remarks on assumption (A1) of a previous version and his suggestions to simplify several proofs have led to a considerable improvement of the paper.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*. 13, 183-198.
- BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*. 3, 141-153.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.
- LEMÂÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*. 13, 199-207.
- NIEUWENBROEK, N.J. (1997). General regression estimator in Bascula: Theoretical background. Research paper no. 9737, Statistics Netherlands.
- NIEUWENBROEK, N.J., and VAN DER VALK, J. (1996). Research paper no. 9629, Statistics Netherlands.
- RAO, C.R. (1973). *Linear Statistical Inference And Its Applications* (2nd edition). New York: John Wiley & Sons, Inc.
- RENSSEN, R.H., NIEUWENBROEK, N.J. and SLOOTBEEK, G. T. (1997). Variance module in Bascula: Theoretical background. Research paper no. 9712, Statistics Netherlands.
- SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J.H. (1992). *Model-assisted Survey Sampling*. New York, Springer-Verlag.
- SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.
- SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees during 2002. An asterisk indicates that the person served more than once.

- | | |
|--|---|
| A. Alavi, <i>Statistics Canada</i> | * P. Lavallée, <i>Statistics Canada</i> |
| C. Alexander, <i>U.S. Bureau of the Census</i> | S. Linacre, <i>Official National Statistics</i> |
| * J-F. Beaumont, <i>Statistics Canada</i> | * S. Lohr, <i>Arizona State University</i> |
| * D.R. Bellhouse, <i>University of Western Ontario</i> | G. Maiti, <i>Iowa State University</i> |
| * D.A. Binder, <i>Statistique Canada</i> | * H. Mantel, <i>Statistics Canada</i> |
| * G.J. Brackstone, <i>Statistics Canada</i> | G. Meeden, <i>University of Minnesota</i> |
| K. Brewer, <i>Australian National University</i> | J.-P. Morin, <i>Statistics Canada</i> |
| J.M. Brick, <i>Westat, Inc.</i> | D. Norris, <i>Statistics Canada</i> |
| T. Buskirk, <i>University of Nebraska-Lincoln</i> | D. Paton, <i>Statistics Canada</i> |
| R. Chambers, <i>University of Southampton</i> | C.R. Perry, <i>NASS</i> |
| M. J. Cho, <i>United States Bureau of Labor Statistics</i> | D. Pfeffermann, <i>Hebrew University</i> |
| C. Clark, <i>U.S. Bureau of the Census</i> | T.E. Raghunathan, <i>University of Michigan</i> |
| M. Cohen, <i>U.S. Bureau of Transportation Statistics</i> | J.N.K. Rao, <i>Carleton University</i> |
| M. Cruddas, <i>Office of National Statistics UK</i> | T.J. Rao, <i>Indian Statistical Institute</i> |
| G. Datta, <i>University of Georgia</i> | E. Rancourt, <i>Statistics Canada</i> |
| * P. Dick, <i>Statistics Canada</i> | J. Reiter, <i>Duke University</i> |
| J. Dumais, <i>Statistics Canada</i> | * L.-P. Rivest, <i>Université Laval</i> |
| J. Eltinge, <i>U.S. Bureau of Labor Statistics</i> | S. Roehrig, <i>Carnegie - Mellon University</i> |
| L. Ernest, <i>Bureau of the Labour Statistics</i> | P.-A. Salamin, <i>Swiss Federal Statistical Office</i> |
| S. Fienberg, <i>Carnegie - Mellon University</i> | N. Schenker, <i>National Center for Health Statistics</i> |
| J.-M. Fillion, <i>Statistics Canada</i> | F.J. Scheuren, <i>National Opinion Research Center</i> |
| W.A. Fuller, <i>Iowa State University</i> | I. Şchiopu-Kratina, <i>Statistics Canada</i> |
| * J. Gambino, <i>Statistics Canada</i> | * R. Sitter, <i>Simon Fraser University</i> |
| * M. Ghosh, <i>University of Florida</i> | C.J. Skinner, <i>University of Southampton</i> |
| G. Glonek, <i>Flinders University</i> | D.D. Smith, <i>U.S. Bureau of the Census</i> |
| J. Green, <i>Westat, Inc.</i> | * E. Stasny, <i>Ohio State University</i> |
| S. Heeringa, <i>University of Michigan</i> | * A. Théberge, <i>Statistics Canada</i> |
| * M.A. Hidioglou, <i>Statistics Canada</i> | Y. Thibaudeau, <i>U.S. Bureau of the Census</i> |
| H. Hogan, <i>U.S. Bureau of Labor Statistics</i> | R.C. Tiwari, <i>National Cancer Institute</i> |
| * D. Judkins, <i>Westat, Inc.</i> | J. Tourigny, <i>Statistics Canada</i> |
| * G. Kalton, <i>Westat, Inc.</i> | * R. Valliant, <i>Westat, Inc.</i> |
| D. Kostanich, <i>U.S. Bureau of the Census</i> | J. Waksberg, <i>Westat, Inc.</i> |
| * P. Kott, <i>NASS</i> | W.E. Winkler, <i>U.S. Bureau of the Census</i> |
| * M. Kovačević, <i>Statistics Canada</i> | K. Wolter, <i>National Opinion Research Center</i> |
| J. Kovar, <i>Statistics Canada</i> | P. Wong, <i>Statistics Canada</i> |
| P. Lahiri, <i>Joint Program in Survey Methodology</i> | Y. You, <i>Statistics Canada</i> |
| * M.D. Larsen, <i>University of Chicago</i> | * W. Yung, <i>Statistics Canada</i> |
| M. Latouche, <i>Statistics Canada</i> | A. Zaslavsky, <i>Harvard University</i> |

Acknowledgements are also due to those who assisted during the production of the 2002 issues: H. Laplante (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Cousineau, C. Ethier, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 18, No. 2, 2002

The Tenth Morris Hansen Lecture Opening Remarks Cynthia Clark	125
Models In the Practice of Survey Sampling (Revisited) Graham Kalton	129
Discussion Chris Skinner	155
Discussion William R. Bell	157
The Eleventh Morris Hansen Lecture Opening Remarks Joseph Waksberg	163
Election Night Estimation Warren J. Mitofsky and Murray Edelman	165
Discussion Martin Frankel	181
Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications Lawrence R. Ernst and Steven P. Paben	185
Two-Phase List-Assisted RDD Sampling J.Michael Brick, David Judkins, Jill Montaquila, and David Morganstein	203
A Design-based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys. An Application in the Dutch Labour Force Survey Jan A. van den Brakel and C.A.M. van Berkel	217
The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling Victor M. Estevao and Carl-Erik Särndal	233
A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality Nojin Kwak and Barry Radler	257
Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets Akimichi Takemura	275
Multiple – Objective Optimal Designs for the Hierarchical Linear Model Mirjam Moerbeek and Weng Kee Wong	291
Book and Software Reviews	305
In Other Journals	317

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents Volume 18, No. 3, 2002

Preface	323
How Best to Hand Out Money: Issues in the Design and Structure of Intergovernmental Aid Formulas Thomas A. Downes and Thomas F. Pogue	329
The Legislative Process and the Use of Indicators in Formula Allocations Dan Melnick	353
Interactions Between Survey Estimates and Federal Funding Formulas Alan M. Zaslavsky and Allen L. Schirm	371
The Canadian Equalization Program Michelle Taylor, Sean Keenan, and Jean-Francois Carboneau	393
Using Survey Data to Allocate Federal Funds for the State Children's Health Insurance Program (SCHIP) John L. Czajka and Thomas B. Jabine	409
Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Funding Formula Evolution Dawn K. Aldridge	429
Impact of Title I Factors on School Year 2000-2001 State Allocations Paul Sanders Brown	441
Federal Formula Allocation for Schools: Historical Perspective and Lessons from New York State James A. Kadamus	465
A Study on the Formulation of an Assessment Scale Methodology: The United Nations Experience in Allocating Budget Expenditures Among Member States Felizardo B. Suzara	481

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

Gemai CHEN, Richard A. LOCKHART & Michael A. STEPHENS Box – Cox transformations in linear models: large sample theory and tests of normality	177
Discussion:	
Comment 1: Karim Maher ABADIR	210
Comment 2: Nancy REID	211
Comment 3: Peter McCULLAGH	212
Comment 4: Peter J. BICKEL	214
Comment 5: Richard A. JOHNSON & Kjell A. DOKSUM	215
Comment 6: Peter M. HOOPER	220
Comment 7: Zhenlin YANG	222
Rejoinder:	
Gemai CHEN, Richard A. LOCKHART & Michael A. STEPHENS	226
Zhenlin YANG	
Median estimation through a regression transformation	235
Donald L. McLEISH	
Highs and lows: some properties of the extremes of a diffusion and applications in finance	243
Hemant ISHWARAN & Mahmoud ZAREPOUR	
Exact and approximate sum representations for the Dirichlet process	269
Antonio CUEVAS, Manuel FEBRERO & Ricardo FRAIMAN	
Linear functional regression: the case of fixed design and functional response	285
Alwell J. OYET	
Minimax A – and D –optimal integer-valued wavelet designs for estimation	301
Christopher A. CAROLAN	
The least concave majorant of the empirical distribution function	317
Christian GENEST & Mireille GUAY	
Worldwide research output in probability and statistics: an update	329
Volume 31 (2003): Subscription rates/Frais d'abonnement	343
Forthcoming Papers/Articles à paraître	346

CONTENTS

TABLE DES MATIÈRES

Volume 30, No. 3, September/septembre 2002, 347-492

Feifang HU & James V. ZIDEK The weighted likelihood	347
Yolanda MUÑOZ MALDONADO, Joan G. STANISWALIS, Louis N. IRWIN & Donna BYERS A similarity analysis of curves	373
Edward SUSKO & Robert NADON Estimation of a residual distribution with small numbers of repeated measurements	383
Sanjoy SINHA & Douglas P. WIENS Minimax weights for generalised M-estimation in biased regression models	401
Daniel B. HALL & Kenneth S. BERENHAUT Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models	415
Yong YOU & J.N.K. RAO A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights	431
Christian GENEST, Jean-François QUESSY & Bruno RÉMILLARD Tests of serial independence based on Kendall's process	441
Paul GUSTAFSON On the simultaneous effects of model misspecification and errors in variables	463
Yanqing SUN, Sufang CUI & Ram C. TIWARI Goodness-of-fit tests for parametric models based on biased samples	475
Forthcoming Papers/Articles à paraître	491
Volume 31 (2003): Subscription rates/Frais d'abonnement	492

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préféablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.

Exemple: Cochran (1977, p. 164).
La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Feifang HU & James V. ZIDEK The weighted likelihood	347
Volanda MUÑOZ MALDONADO, Joan G. STANISWALIS, Louis N. IRWIN & Donna BYERS A similarity analysis of curves	373
Edward SUSKO & Robert NADON Estimation of a residual distribution with small numbers of repeated measurements	383
Sanjoy SINHA & Douglas P. WIENS Minimax weights for generalised M-estimation in biased regression models	401
Daniel B. HALL & Kenneth S. BERENHAUT Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models	415
Yong YOU & J.N.K. RAO A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights	431
Christian GENEST, Jean-François QUESY & Bruno RÉMILLARD Tests of serial independence based on Kendall's process	441
Paul GUSTAFSON On the simultaneous effects of model misspecification and errors in variables	463
Yangting SUN, Surang CUI & Ram C. TIWARI Goodness-of-fit tests for parametric models based on biased samples	475
Forthcoming Papers/Articles à paraître	491
Volume 31 (2003): Subscription rates/Frais d'abonnement	492

Volume 30, No. 2, June/juin 2002, 177-346

Gemai CHEN, Richard A. LOCKHART & Michael A. STEPHENS	Box - Cox transformations in linear models: large sample theory and tests of normality	177
Discussion:		
Comment 1: Karim Maher ABADIR	Comment 2: Nancy REID	210
Comment 3: Peter McCULLAGH	Comment 4: Peter J. BICKEL	212
Comment 5: Richard A. JOHNSON & Kjell A. DOKSUM	Comment 6: Peter M. HOOPER	215
Comment 7: Zhenlin YANG		222
Rejoinder:		
Gemai CHEN, Richard A. LOCKHART & Michael A. STEPHENS		226
Zhenlin YANG	Median estimation through a regression transformation	235
Donald L. McLEISH	Highs and lows: some properties of the extremes of a diffusion and applications in finance	243
Hemant ISHWARAN & Mahmoud ZAREPOUR	Exact and approximate sum representations for the Dirichlet process	269
Antonio CUEVAS, Manuel FEBREIRO & Ricardo FRAIMAN	Linear functional regression: the case of fixed design and functional response	285
Alwell J. OYEY	Minimax A - and D - optimal integer-valued wavelet designs for estimation	301
Christopher A. CAROLAN	The least concave majorant of the empirical distribution function	317
Christian GENEST & Mireille GUY	Worldwide research output in probability and statistics: an update	329
Volume 31 (2003): Subscription rates/Frais d'abonnement		343
Forthcoming Papers/Articles à paraître		346

Contents
Volume 18, No. 3, 2002

323	Preface
329	How Best to Hand Out Money: Issues in the Design and Structure of Intergovernmental Aid Formulas Thomas A. Downes and Thomas F. Pogue
353	The Legislative Process and the Use of Indicators in Formula Allocations Dan Melnick
371	Interactions Between Survey Estimates and Federal Funding Formulas Alan M. Zaslavsky and Allen L. Schirm
393	The Canadian Equalization Program Michelle Taylor, Sean Keenan, and Jean-Francois Carbonneau
409	Using Survey Data to Allocate Federal Funds for the State Children's Health Insurance Program (SCHIP) John L. Czajka and Thomas B. Jabine
429	Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Funding Formula Evolution Dawn K. Aldridge
441	Impact of Title I Factors on School Year 2000-2001 State Allocations Paul Sanders Brown
465	Federal Formula Allocation for Schools: Historical Perspective and Lessons from New York State James A. Kadamus
481	A Study on the Formulation of an Assessment Scale Methodology: The United Nations Experience in Allocating Budget Expenditures Among Member States Felizardo B. Suzara

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

Contents Volume 18, No. 2, 2002

The Tenth Morris Hansen Lecture Opening Remarks Cynthia Clark	125
Models In the Practice of Survey Sampling (Revised) Graham Kalton	129
Discussion Chris Skinner	155
Discussion William R. Bell	157
The Eleventh Morris Hansen Lecture Opening Remarks Joseph Waksberg	163
Election Night Estimation Warren J. Mitofsky and Murray Edelman	165
Discussion Martin Frankel	181
Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications Lawrence R. Ernst and Steven P. Faben	185
Two-Phase List-Assisted RDD Sampling J. Michael Brick, David Judkins, Jill Montaquila, and David Morganstein	203
A Design-based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys. Jan A. van den Brakel and C.A.M. van Berkel	217
The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling Victor M. Estévez and Carl-Erik Särndal	233
A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality Nojin Kwak and Barry Kader	257
Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets Akimichi Takemura	275
Multiple – Objective Optimal Designs for the Hierarchical Linear Model Mirjam Moerbeek and Weng Kee Wong	291
Book and Software Reviews	305
In Other Journals	317

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article durant l'année 2002. Un astérisque indique que la personne a participé plus d'une fois.

- * A. Alavi, *Statistique Canada*
- * C. Alexander, *U.S. Bureau of the Census*
- * J.-F. Beaumont, *Statistique Canada*
- * D.R. Bellhouse, *University of Western Ontario*
- * D.A. Binder, *Statistique Canada*
- * G.J. Brackstone, *Statistique Canada*
- K. Brewer, *Australian National University*
- J.M. Brick, *Westat, Inc.*
- T. Buskirk, *University of Nebraska-Lincoln*
- R. Chambers, *University of Southampton*
- M. J. Cho, *United States Bureau of Labor Statistics*
- C. Clark, *U.S. Bureau of the Census*
- M. Cohen, *U.S. Bureau of Transportation Statistics*
- M. Cruddas, *Office of National Statistics UK*
- G. Datta, *University of Georgia*
- * P. Dick, *Statistique Canada*
- J. Dunnais, *Statistique Canada*
- J. Eitinge, *U.S. Bureau of Labor Statistics*
- L. Ernest, *Bureau of the Labour Statistics*
- S. Fienberg, *Carnegie - Mellon University*
- J.-M. Fillion, *Statistique Canada*
- W.A. Fuller, *Iowa State University*
- * J. Gambino, *Statistique Canada*
- * M. Ghosh, *University of Florida*
- G. Glonek, *Flinders University*
- J. Green, *Westat, Inc.*
- * S. Heeringa, *University of Michigan*
- * M.A. Hidiroglou, *Statistique Canada*
- H. Hogan, *U.S. Bureau of Labor Statistics*
- * G. Kalton, *Westat, Inc.*
- D. Kostanich, *U.S. Bureau of the Census*
- * P. Kott, *NASS*
- * M. Kovachević, *Statistique Canada*
- J. Kovar, *Statistique Canada*
- P. Lahiri, *Joint Program in Survey Methodology*
- * M.D. Larsen, *University of Chicago*
- M. Latouche, *Statistique Canada*
- * P. Lavallée, *Statistique Canada*
- S. Linacre, *Official National Statistics*
- * S. Lohr, *Arizona State University*
- G. Mailt, *Iowa State University*
- * H. Mantel, *Statistique Canada*
- J.-P. Morin, *Statistique Canada*
- D. Norris, *Statistique Canada*
- D. Paton, *Statistique Canada*
- C.R. Perry, *NASS*
- D. Pfeffermann, *Hebrew University*
- T.E. Raghunathan, *University of Michigan*
- J.N.K. Rao, *Carleton University*
- T.J. Rao, *Indian Statistical Institute*
- E. Rancourt, *Statistique Canada*
- J. Reiter, *Duke University*
- * L.-P. Rivest, *Université Laval*
- S. Roehrig, *Carnegie - Mellon University*
- P.-A. Salamin, *Swiss Federal Statistical Office*
- N. Schenker, *National Center for Health Statistics*
- F.J. Scheuren, *National Opinion Research Center*
- I. Schiopu-Krautha, *Statistique Canada*
- * R. Sitter, *Simon Fraser University*
- C.J. Skinner, *University of Southampton*
- D.D. Smith, *U.S. Bureau of the Census*
- * E. Stasny, *Ohio State University*
- * A. Theberge, *Statistique Canada*
- Y. Thibaudau, *U.S. Bureau of the Census*
- R.C. Tiwari, *National Cancer Institute*
- J. Tourigny, *Statistique Canada*
- * R. Vallian, *Westat, Inc.*
- J. Waksberg, *Westat, Inc.*
- W.E. Winkler, *U.S. Bureau of the Census*
- K. Wolter, *National Opinion Research Center*
- P. Wong, *Statistique Canada*
- * Y. You, *Statistique Canada*
- W. Yung, *Statistique Canada*
- A. Zaslawsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2002: H. Laplante (Division de la diffusion) et L. Perteault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à C. Cousineau, C. Ethier et D. Lemire de la Division des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

Les opinions exprimées dans l'article n'engagent que les auteurs et ne reflètent pas forcément les politiques de Statistics Netherlands.

BIBLIOGRAPHIE

- ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.
- BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

- NIEUWENBROEK, N.J. (1997). General regression estimator in Bascula: Theoretical background. Article de recherche, numéro 9737, Statistics Netherlands.
- NIEUWENBROEK, N.J., et VAN DER VALK, J. (1996). Article de recherche, numéro 9629, Statistics Netherlands.
- RAO, C.R. (1973). *Linear Statistical Inference And Its Applications* (2^e édition). New York: John Wiley & Sons, Inc.
- RENSSEN, R.H., NIEUWENBROEK, N.J., et SLOOTBEEK, G. T. (1997). Variance module in Bascula: Theoretical background. Article de recherche, numéro 9712, Statistics Netherlands.
- SÄRNDAAL, C.E., SWENSSON, B. et WRETMAN, J.H. (1992). *Model-assisted Survey Sampling*. New York, Springer-Verlag.
- SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.
- SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Si \mathbf{X}^S est de rang $r > d$, alors une application de (3) donnera r éléments diagonaux non nuls et $d - r$ éléments diagonaux nuls. Si nous trouvons un élément diagonal nul, alors nous fixons à zéro la valeur des éléments de la ligne et de la colonne qui lui correspondent. Subséquentement, par échange élémentaire de ligne et de colonne, nous obtenons la matrice triangulaire supérieure suivante :

$$\cdot \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{pmatrix} = \mathbf{U}$$

colonnes, nous échangeons aussi les éléments de X_j et

où, par construction, X_{1s} est de plein rang et E est une matrice non singulière d'ordre $p \times p$. Mais, puisque

est une inverse généralisée de $(X_{1S} X_{2S})' V_S (X_{1S} X_{2S})$, nous savons que $G_S = E' G_S E$ est un inverse généralisé de $X_S' V_S X_S$. L'insertion de cet inverse généralisé dans

$$\begin{pmatrix} \text{JH}^2 \mathbf{x} - \text{x}_2^1 \mathbf{1} \\ \text{JH}^1 \mathbf{x} - \text{x}_1^1 \mathbf{1} \end{pmatrix} \mathbf{G}_{S'}^k(\mathbf{x}_1^k, \mathbf{x}_2^k) \mathbf{v}^k + \mathbf{1}^k \mathbf{w}^k = \mathbf{w}^k$$

5. L'ÉNOUËTE SUR LA POPULATION ACTIVE

DES PAYS-BAS

DES PAYS-BAS

6. REMERCIEMENTS

constater facilement en prenant $\mathbf{v} = \mathbf{w} - \mathbf{d}_s$ avec $\mathbf{d}_s = (\pi_1^{-1}, \dots, \pi_n^{-1})'$. Nous pouvons montrer comme suit l'invariance des coefficients de régression en fonction du choix de \mathbf{G}_s et, donc, l'invariance de l'estimateur de régression généralisée. Posons que \mathbf{F}_s est un autre inverse généralisé de $\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s$, différent de \mathbf{G}_s . Alors, nous

$$\begin{aligned} \mathbf{X}_s \mathbf{G}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) &= \mathbf{X}_s \mathbf{G}_s \mathbf{X}_s' \mathbf{v} & \text{d'après (H1)} \\ \mathbf{X}_s \mathbf{F}_s \mathbf{X}_s' \mathbf{v} &= \mathbf{X}_s \mathbf{F}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) & \text{d'après (P3)} \\ \mathbf{X}_s \mathbf{F}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) &= \mathbf{X}_s \mathbf{F}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) & \text{d'après (H1)} \end{aligned}$$

Donc, il est vérifié que $\mathbf{x}_k' \mathbf{G}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}})$ ne varie pas en fonction de \mathbf{G}_s pour tout $k \in S$, ce qui implique que les coefficients de régression ne dépendent pas du choix de \mathbf{G}_s . Le fait que ces coefficients de régression reproduisent les totaux de population des variables auxiliaires découle de la série d'équations suivantes :

$$\begin{aligned} \sum_{k \in S} \mathbf{w}_k \mathbf{x}_k &= \mathbf{x}_{\text{HT}} + \sum_{k \in S} \mathbf{x}_k \mathbf{x}_k' \mathbf{G}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) \\ &= \mathbf{x}_{\text{HT}} + (\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s) \mathbf{G}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) \\ &= \mathbf{x}_{\text{HT}} + (\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s) \mathbf{G}_s \mathbf{X}_s' \mathbf{v} \text{ d'après (H1)} \\ &= \mathbf{x}_{\text{HT}} + \mathbf{X}_s' \mathbf{v} \text{ d'après (P2) et (P4)} \\ &= \mathbf{x}_{\text{HT}} + (\mathbf{I}_x - \mathbf{x}_{\text{HT}}) = \mathbf{I}_x \text{ d'après (H1)} \end{aligned}$$

Pour terminer cette section, examinons de plus près l'hypothèse énoncée pour certains modèles de pondération bien connus. Dans le cas de la stratification, l'interprétation de pondération est décrit par un croisement complet de variables nominales. L'interprétation de (H1) est simple. Plus précisément, (H1) est satisfaite si, l'échantillon correspondent à des strates à posteriori vides dans l'échantillon. Considérons maintenant la stratification a posteriori incomplète où le modèle de pondération comprend plusieurs termes, chacun décrivant un croisement complet de variables nominales et correspondant donc, chacun, à une stratification a posteriori. Alors, une condition nécessaire pour que (H1) soit satisfaite est que les strates à posteriori vides dans la population correspondent à des strates vides dans la population pour se produire lorsque nous essayons de procéder au calage sur un certain nombre de croisements complets plus grands que la taille de l'échantillon.

L'hypothèse n'est pas aussi simple dans le cas de la pondération cohérente entre les personnes et les ménages (voir, par exemple, Lemaître et Dufour 1987). Cette situation est due à la redéfinition de la variable auxiliaire. Par exemple, si \mathbf{x}_k est une variable définie au niveau de la personne et que, à partir de cette variable, on en définit une

4. CALCUL DES COEFFICIENTS DE RÉGRESSION DANS BASCULA

À la section précédente, nous avons montré que les coefficients de régression généralisée $\mathbf{w}_k = \pi_k^{-1} + \lambda_k \mathbf{x}_k' \mathbf{G}_s (\mathbf{I}_x - \mathbf{x}_{\text{HT}})$ ne variaient pas en fonction du choix de \mathbf{G}_s . À la présente section, nous montrons comment calculer ces coefficients. Pour cela, nous commençons par la décomposition de Cholesky de la matrice (semi) définie positive $\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s$, (voir Söber 1977, page 322). Si \mathbf{X}_s est de plein rang, alors $\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s$ est une matrice définie positive qui peut être exprimée de façon unique sous la forme $\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s = \mathbf{U}' \mathbf{U}$, où \mathbf{U} est une matrice triangulaire supérieure à éléments diagonaux positifs. Représentons par a_{ij} le $ij^{\text{ème}}$ élément de $\mathbf{X}_s' \mathbf{A}_s \mathbf{X}_s$. Alors, \mathbf{U} peut être calculée, ligne par ligne, selon

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2} \quad \text{pour } i = 1, \dots, p$$

$$d_{ij} = \frac{a_{ij}}{\sum_{k=1}^{i-1} u_{ki} u_{kj}} \quad \text{pour } j = i + 1, \dots, p.$$

(3)

nouvelle au niveau du ménage, disons \mathbf{z}_k , alors (H1) devrait être définie en fonction de $\mathbf{Z}_s = (\mathbf{z}_1', \dots, \mathbf{z}_n')$ plutôt que \mathbf{X}_s . Autrement dit, (H1) est satisfaite s'il existe un vecteur \mathbf{n} des pondérations, représenté par \mathbf{w} , tel que $\mathbf{Z}_s' \mathbf{w} = \mathbf{I}_x$. Dans de nombreuses situations (ordinaires), la variable linéaire couverte par \mathbf{Z}_s coïncidera avec la variable linéaire couverte par \mathbf{X}_s . Dans ces situations, la méthode de Lemaître et Dufour n'influe pas sur la validité de (H1). Cependant, il pourrait ne pas en être ainsi dans certains cas. L'exemple simplifié qui suit en donne une illustration. Soit \mathbf{x}_k , le sexe de la $k^{\text{ème}}$ personne, disons $\mathbf{x}_k = (0, 1)'$ si cette personne est une femme et $\mathbf{x}_k = (1, 0)'$ si elle est un homme. Conformément à la méthode de Lemaître et Dufour (1987), représentons par \mathbf{z}_k la $k^{\text{ème}}$ moyenne des ménages pour \mathbf{x}_k lorsque k appartient au $j^{\text{ème}}$ ménage. En outre, posons que la population contient N_1 hommes et N_2 femmes, et que l'on tire un échantillon de dix ménages. Supposons que chaque ménage échantillonné comprend deux personnes, à savoir un homme et une femme. Ceci donne $\mathbf{z}_k = (1/2, 1/2)'$ pour tout $k \in S$. Dans cet exemple, la variable linéaire couverte par \mathbf{Z}_s est un sous-espace linéaire de la variété linéaire couverte par \mathbf{X}_s . Si $N_1 = N_2$, alors (H1) est satisfaite. Sinon, si $N_1 \neq N_2$, (H1) n'est pas satisfaite. Plus précisément, lorsque la méthode de Lemaître et Dufour est appliquée à un modèle de pondération assez grand, la variété linéaire couverte par \mathbf{Z}_s pourrait être un sous-espace approprié de la variété linéaire couverte par \mathbf{X}_s . Alors, (H1) n'est satisfaite que si \mathbf{x}_k appartient accidentellement à ce sous-espace.

les matrices diagonales sont strictement positives et X est une matrice de plan d'expérience d'ordre $n \times p$ qui résulte du modèle de pondération. Pour une discussion détaillée des matrices inverses généralisées, consulter Searle (1971) et Rao (1973).

Avant d'énoncer ces propriétés, nous passons brièvement en revue la définition d'un inverse généralisé. Considérons une matrice A d'ordre $p \times q$ de n importe quel rang et représentons par $Ax = y$ un système d'équations cohérentes; autrement dit, toute relation linéaire entre les lignes de A existe aussi entre les éléments correspondants de y . Une inverse généralisée de A est une matrice A^- d'ordre $q \times p$ telle que $x = A^-y$ est une solution de ce système d'équations. Il est facile de vérifier que l'existence de A^- implique $AA^-A = A$ (choisir y comme étant la i -ième colonne de A). Inversement, si A^- satisfait $AA^-A = A$ et que $Ax = y$ est cohérent, alors $A(A^-y) = A(A^-Ax) = Ax = y$ et, donc, A^-y est une solution. Par conséquent, une autre définition est qu'une matrice inverse généralisée de A est toute matrice A^- telle que $AA^-A = A$.

Maintenant, si G représente un inverse généralisé de $X'AX$, alors les propriétés suivantes de G sont

prouvées dans Searle (1971) pour $A = I_n$:

- (P1) G' est également un inverse généralisé de $X'AX$,
- (P2) $XXGX'AX = X$, autrement dit $GX'AX$ est un inverse généralisé de X ,
- (P3) $XXGX'$ ne dépend pas du choix de G ,
- (P4) $XXGX' = XXG'X'$, que G soit symétrique ou non.

Les preuves de (P1) à (P4) pour une matrice diagonale sont presque identiques à celles de Searle (1971, chapitre 1.5, théorème 7) et ne sont donc pas reproduites ici.

3. L'ESTIMATEUR DE RÉGRESSION GÉNÉRALISÉE

Considérons une population finie U de N unités à partir de laquelle on tire un échantillon S de n unités sans remise. Représentons par π_k la probabilité d'inclusion de premier ordre de la k -ième unité $k = 1, \dots, N$. Nous associons à chaque unité un vecteur de variables étudiées y^k . Alors, la matrice de données pour les unités échantillonnées est représentée par $Y_S = (y^1, \dots, y^n)'$. Nous faisons la distinction entre les variables étudiées pour lesquelles on connaît les totaux de population (variables auxiliaires) et celles pour lesquelles on ne connaît pas ces totaux. Le point de départ de la définition d'un estimateur de régression généralisée (Särndal et coll. 1992) est la spécification du modèle de pondération, autrement dit le choix de l'ensemble de variables auxiliaires qu'il faut utiliser dans l'estimation. En représentant cet ensemble par la matrice $X_S = (x^1, \dots, x^n)'$ d'ordre $n \times p$ le nom de matrice de plan d'expérience, qui est, par définition, un sous-ensemble de colonnes de Y_S . Le

vecteur des totaux connus de population de x est représenté par $x_{HT} = \sum_{k \in S} \pi_k x^k$, l'estimateur d'Horvitz-Thompson pour x ; alors, étant donné x , l'estimateur de régression généralisée du vecteur des totaux de population de la i -ième variables étudiées y^k est défini par

$$(1) \quad \hat{t}_{(i)}^{grég} = y_{(i)}^{HT} + B_{(i)}'(t^x - x^{HT})$$

$$B = G_S'X_S'V_SX_S$$

En fonction des coefficients de régression, cet estimateur de régression généralisée peut aussi s'écrire sous la forme

$$(2) \quad \hat{t}_{(i)}^{grég} = \sum_{k \in S} w_k^{(i)} y_k^{(i)}$$

$$w_k = \pi_k^{-1} + \lambda_k x_k' G_S(t^x - x^{HT}).$$

Ici, G_S représente une inverse généralisée de $X_S'V_SX_S$ et $V_S = \text{diag}(v_1, \dots, v_n)$ est une matrice diagonale ne contenant strictement que des entées positives.

Comme le modèle de pondération, la matrice diagonale V_S doit être spécifiée par l'utilisateur. Souvent, on prend $V_S = \Pi_S^{-1}$, où $\Pi_S = \text{diag}(\pi_1, \dots, \pi_n)$ et $\sum_S = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ où σ_k^2 est interprété comme étant la variance de variables aléatoires indépendantes dont certaines variables étudiées sont censées être le résultat conformément à un certain modèle de superpopulation (voir Särndal et coll. 1992). Il est nécessaire de connaître tous les σ_k^2 jusqu'à un facteur d'échelle commun. Un cas spécial important est $\sigma_k^2 = \sigma^2$, c'est-à-dire celui où les variances modélisées sont toutes les mêmes. On obtient alors l'estimateur de régression proposé par Bethlehem et Keller (1987). Si les unités de population représentent des ménages (de taille m_k) et que nous prenons $\sigma_k^2 = m_k \sigma^2$, nous arrivons à l'estimateur proposé par Lemaitre et Dufour (1987) pour obtenir des coefficients de pondération cohérents entre les personnes et les ménages. Examinant le problème sous un autre angle, Alexander (1987) a procédé à l'estimation par les moindres carrés généralisés qui aboutit essentiellement au même estimateur.

Plus bas, nous montrons que les coefficients de régression ne varient pas en fonction du choix de G_S . À cette fin, nous émettons l'hypothèse suivante :

$$(H1) \quad \text{il existe un vecteur } n \text{ des pondérations, représentant par } w, \text{ tel que } X_S'w = t^x.$$

Clairement, cette hypothèse dit que $X_S'w = t^x$ est un système d'équations cohérentes. Il est intéressant de noter que ce système correspond précisément à l'ensemble d'équations de calage si l'on considère l'estimateur de régression généralisée comme un cas spécial de l'estimateur par calage (voir, par exemple, Deville et Särndal 1992). Si $X_S'w = t^x$ est un système d'équations cohérentes, alors il en est de même de $X_S'v = (t^x - x^{HT})$. On peut le

De l'utilisation des matrices inverses généralisées dans la théorie de l'échantillonnage

ROBERT H. RENSSSEN et GERARD H. MARTINUS¹

RÉSUMÉ

En théorie, il est coutumier de définir les estimateurs de régression généralisée au moyen de modèles de pondération de plein rang; autrement dit, la matrice de plan d'expérience qui correspond au modèle de pondération est de plein rang. Il est bien connu que, pour de tels modèles de pondération, les coefficients de régression généralisée reproduisent les totaux (communs) de population des variables auxiliaires incluses dans le modèle. Toutefois, en pratique, il arrive souvent que le modèle de pondération ne soit pas de plein rang, particulièrement s'il est établi pour une stratification à posteriori incomplète. Au moyen de la théorie des matrices inverses généralisées, nous montrons dans quelles circonstances cette propriété de cohérence demeure valide. À titre d'exemple non trivial, nous discutons de la pondération cohérente entre les personnes et les ménages proposée par Lemaitre et Dufour (1987). Puis, nous montrons comment la théorie est appliquée dans le logiciel Bascula.

MOTS CLÉS : Bascula; estimateur de régression généralisée; pondération.

1. INTRODUCTION

Lors d'enquêtes par sondage, on se sert couramment de méthodes de pondération basées sur l'estimateur de régression généralisée pour faire les corrections pour tenir compte de l'erreur d'échantillonnage ainsi que de l'erreur non due à l'échantillonnage; consulter, par exemple, Bethlehem et Keller (1987) et Sæmstad, Swensson et Westman (1992). Cependant, l'une des complications de l'utilisation des estimateurs de régression généralisée tient au fait que nombre de modèles de pondération sont fondés sur une stratification à posteriori incomplète, ce qui produit des matrices de plan d'expérience qui ne sont pas de plein rang. Habituellement, on résout ce problème en utilisant une matrice réduite de plan d'expérience, que l'on peut rajuster correctement les totaux de population. Souvent, il est assez facile de repérer d'avance les redondances, d'après les spécifications du modèle de pondération. Cependant, dans le cas de certains modèles, le dépistage de ces redondances n'est pas pratique.

Par exemple, supposons que nous ayons affaire à une stratification à posteriori fondée sur le croisement complet entre deux variables nominales A et B , pour lesquelles on connaît les dénombrements de population pour chaque cellule. Les dénombrements échantillonnables pourraient être faibles ou nuls pour certaines cellules. Le cas échéant, nous pouvons obtenir de nouvelles classifications, A' à partir de A et B' à partir de B , en fusionnant des catégories et de définir le schéma plus parcimonieux suivant : $A' + B' + A' \times B'$. Partant de cette stratification à posteriori incomplète, nous procédons au calage simultané sur trois ensembles de dénombrements, à savoir les dénombrements marginaux de A , les dénombrements marginaux de B et les

Nous nous intéressons principalement à l'utilisation des inverses généralisées dans le cadre de l'estimateur de régression généralisée. Donc, nous donnons uniquement certaines propriétés d'une inverse généralisée de la forme $X' \Lambda X$, où Λ est une matrice diagonale d'ordre $n \times n$ dont

2. MATRICES INVERSES GÉNÉRALISÉES

L'Enquête sur la population active des Pays-Bas. Nous discutons brièvement du modèle de pondération de des coefficients de régression. Enfin, à la section 5, nous appliquons dans Bascula ajouter la référence pour le calcul des ménages. À la section 4, nous décrivons l'algorithme qui est incomplet et la pondération cohérente entre personnes et connus, comme la pondération par stratification à posteriori régularisée pour certains modèles de pondération bien section 3, nous discutons du respect de cette condition de fonction du choix de l'inverse généralisée. À la fin de la section 3, nous montrons que cet estimateur ne varie pas en 1992), nous montrons que cet estimateur ne varie pas en contexte d'estimation par calage (voir Deville et Sæmstad condition de régularité qui s'interprète bien dans un pas nécessairement être de plein rang. Étant donné une généralisée pour les modèles de pondération qui ne doivent section 3, nous définissons l'estimateur de régression propriétés des matrices inverses généralisées. À la section 2, nous décrivons brièvement certaines d'expérience de ce genre.

À la section 2, nous décrivons brièvement certaines d'expérience de ce genre. Nous décrivons la réduction d'une matrice de plan généralisées, de la réduction d'une matrice de plan décrit le contexte théorique, basé sur les matrices inverses d'après les classifications du modèle. Le présent article diffèrents, il est difficile de déterminer les redondances aussi B et B' figurent dans des termes de pondération dénombrements des cellules de $A' \times B'$. Puisque A et A' (et

¹ Robert H. Renssen et Gerard H. Martinus, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands.

- VIHMA, T. (1981). Health hazards and stress factors in small industry-Prevalence study in the province of Uusimaa with special reference to the type of industry and the occupational title as classifications for the description of occupational health problems. *Scandinavian Journal of Work, Environment and Health*, 7, Suppl. 3, 1-149.
- WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

$$E[w(s)] = \sum_{s=1}^S w(s) p(s) = \sum_{s=1}^S \left\{ \sum_{j=1}^J (n_{sj}^t(s) - r_{sj}^t)^2 + \sum_{j=1}^J (n_{sj}^f(s) - r_{sj}^f)^2 \right\} p(s)$$

$$= \sum_{j=1}^J \sum_{s=1}^S (n_{sj}^t(s) - r_{sj}^t)^2 p(s) + \sum_{j=1}^J \sum_{s=1}^S (n_{sj}^f(s) - r_{sj}^f)^2 p(s) = \sum_{j=1}^J \left(\sum_{s=1}^S (r_{sj}^t - \lfloor r_{sj}^t \rfloor) + \sum_{s=1}^S (r_{sj}^f - \lfloor r_{sj}^f \rfloor) \right) \left(1 + \lfloor r_{sj}^t \rfloor - \lfloor r_{sj}^f \rfloor \right)$$

BIBLIOGRAPHIE

- BRYANT, E.C., HARTLEY, H.O. et JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.
- CAUSEY, B.D., COX, L.H. et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- ENGLE, M., MARSDEN, G. et POLLOCK, S.W. (1971). Child work and social class. *Psychiatry*, 34, 140-150.
- GOODMAN, R., et KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- HESS, T., RIEDEL, D.C. et FITZPATRICK, T.B. (1976). *Probability Sampling of Hospitals and Patients*. University of Michigan, Ann Arbor, deuxième édition.
- JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.
- JESSEN, R.J. (1973). Some properties of probability lattice sampling. *Journal of the American Statistical Association*, 68, 20-28.
- JESSEN, R.J. (1975). Square and cubic lattice sampling. *Biometrics*, 31, 449-471.
- KELLY, J.K., GOLDEN, B.L. et ASSAD, A.A. (1993). The controlled rounding problem: complexity and computational experience. *European Journal of Operational Research*, 65, 207-217.
- LAHIRI, P. et MUKERJEE, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistica*, 10, 1171-1178.
- LU, W. (2000). Multi-way stratification by linear programming made practical. Thèse de maîtrise, Simon Fraser University.
- RAO, J.N.K., HARTLEY, H.O. et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Serie B*, 24, 482-491.
- RAO, J.N.K. et NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.
- RAO, J.N.K. et NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.
- SITTER, R.R., et SKINNER, C.J. (1994). Stratification multidimensionnelle par programmation linéaire. *Techniques d'enquête*, 20, 69-78.
- SKINNER, C.J., HOLMES, D.J. et HOLT, D. (1994). Multiple frame sampling for multiple stratification. *International Statistical Review*, 62, 333-347.

5. CONCLUSION

Nous proposons une méthode de stratification à deux dimensions qui étend l'applicabilité de la méthode de programmation linéaire de Sitter et Skinner (1994) à des problèmes de taille nettement plus grande. La méthode se concentre sur la façon de construire un petit ensemble d'échantillons candidats « représentatifs » en utilisant une méthode d'échantillonnage avec probabilités inégales qui génère des échantillons candidats satisfaisant presque les contraintes de RPPA du problème de programmation linéaire, puis sur l'application de la méthode de programmation linéaire à cet ensemble plus petit d'échantillons.

Il convient de souligner que la méthode de programmation linéaire s'étend facilement aux plans d'échantillonnage stratifiés à plusieurs degrés. Puisqu'il n'existe aucune différence fondamentale entre la méthode de programmation linéaire originale et l'extension proposée ici, cette généralisation tient aussi pour la méthode proposée. Dans le même esprit, on peut aussi envisager une discussion des questions ayant trait à l'estimation de la variance des estimateurs résultant de Sitter et Skinner (1994).

Notons toutefois que, lorsqu'on impose des contraintes en fixant autour des n_{ij}^t des bornes telles qu'ils soient des entiers, le problème ressemble à un problème d'arrondissement (voir Kelly, Golden et Assad 1993, et les références qu'ils citent), mais nous n'explorons pas cette question ici.

REMERCIEMENTS

Ces travaux ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE 1

Preuve de (16) : $n_{ij}^t(s) - \lfloor r_{ij}^t \rfloor \sim \text{Bernoulli}(r_{ij}^t - \lfloor r_{ij}^t \rfloor)$ et la variance $(r_{ij}^t - \lfloor r_{ij}^t \rfloor)(1 + \lfloor r_{ij}^t \rfloor - r_{ij}^t)$. Ceci implique que

$$\sum_{s=1}^S (n_{ij}^t(s) - r_{ij}^t)^2 p(s) = E(n_{ij}^t(s) - r_{ij}^t)^2 V(n_{ij}^t(s))$$

$$= V(n_{ij}^t(s) - \lfloor r_{ij}^t \rfloor) \left(1 + \lfloor r_{ij}^t \rfloor - r_{ij}^t \right),$$

et, suivant un argument similaire, que

$$\sum_{s=1}^S (n_{ij}^f(s) - r_{ij}^f)^2 p(s) = (r_{ij}^f - \lfloor r_{ij}^f \rfloor)(1 + \lfloor r_{ij}^f \rfloor - r_{ij}^f).$$

Par conséquent, avec $w(s)$ définie en (3),

sous stratification proportionnelle ($n=151$)

12	0.99	0.63	0.31	0.13	0.23	0.79	0.83	0.96	0.95	0.82	0.81	0.88	0.8	0.73
9	0.40	0.67	0.01	0.77	0.92	0.81	0.54	0.60	0.60	0.22	0.00	0.00	0.17	0.43
9	0.31	0.02	0.60	0.65	0.31	0.42	0.05	0.35	0.35	0.17	0.41	0.50	0.06	0.32
10	0.26	0.02	0.38	0.66	0.42	0.10	0.96	0.82	0.93	0.62	0.81	0.81	0.78	0.12
9	0.85	0.16	0.01	0.05	0.49	0.05	0.79	0.61	0.36	0.35	0.28	0.02	0.34	0.22
9	0.87	0.68	0.55	0.23	0.74	0.38	0.49	0.46	0.12	0.35	0.93	0.35	0.80	0.32
11	0.77	0.14	0.70	0.58	0.74	0.74	0.33	0.74	0.19	0.85	0.48	0.89	0.78	0.47
10	0.75	0.17	0.72	0.04	0.56	0.91	0.27	0.63	0.86	0.96	0.20	0.94	0.30	0.65
10	0.42	0.95	0.20	0.74	0.24	0.11	0.81	0.33	0.33	0.41	0.91	0.82	0.72	0.03
11	0.49	0.78	0.87	0.94	0.47	0.43	0.88	0.71	0.04	0.29	0.20	0.94	0.03	0.39
10	0.76	0.58	0.55	0.26	0.14	0.34	0.95	0.53	0.39	0.37	0.79	0.69	0.88	0.71
11	0.51	0.55	0.82	0.75	0.57	0.64	0.46	0.18	0.55	0.81	0.54	0.81	0.54	0.51
12	0.75	0.94	0.77	0.16	0.60	0.44	0.06	0.30	0.69	0.85	0.56	0.77	0.30	0.72
9	0.53	0.96	0.07	0.71	0.56	0.48	0.13	0.03	0.37	0.51	0.78	0.05	0.25	0.73
9	0.34	0.56	0.7	0.18	0.95	0.06	0.11	0.77	0.82	0.44	0.01	0.44	0.87	0.25
151	9	8	7	6	8	7	9	7	7	8	7	9	7	7

Enquête sur la santé au travail, Viïma (1981)
 Nombres prévus d'unités d'échantillonnage dans les cellules sous
 stratification proportionnelle ($n = 100$)

Type de branche d'activité	Nombre d'employés
Produits alimentaires	5-9
	10-19
	20-49
	50-99
	100-199
	200-499
	500-999
	1000-1999
	2000-4999
	5000-9999
	10000-24999
	25000-49999
	50000-99999
	100000-249999
	250000-499999
	500000-999999
	1000000-2499999
	2500000-4999999
	5000000-9999999
	10000000-24999999
	25000000-49999999
	50000000-99999999
	100000000-249999999
	250000000-499999999
	500000000-999999999
	1000000000-2499999999
	2500000000-4999999999
	5000000000-9999999999
	10000000000-24999999999
	25000000000-49999999999
	50000000000-99999999999
	100000000000-249999999999
	250000000000-499999999999
	500000000000-999999999999
	1000000000000-2499999999999
	2500000000000-4999999999999
	5000000000000-9999999999999
	10000000000000-24999999999999
	25000000000000-49999999999999
	50000000000000-99999999999999
	100000000000000-249999999999999
	250000000000000-499999999999999
	500000000000000-999999999999999
	1000000000000000-2499999999999999
	2500000000000000-4999999999999999
	5000000000000000-9999999999999999
	10000000000000000-24999999999999999
	25000000000000000-49999999999999999
	50000000000000000-99999999999999999
	100000000000000000-249999999999999999
	250000000000000000-499999999999999999
	500000000000000000-999999999999999999
	1000000000000000000-2499999999999999999
	2500000000000000000-4999999999999999999
	5000000000000000000-9999999999999999999
	10000000000000000000-24999999999999999999
	25000000000000000000-49999999999999999999
	50000000000000000000-99999999999999999999
	100000000000000000000-249999999999999999999
	250000000000000000000-499999999999999999999
	500000000000000000000-999999999999999999999
	1000000000000000000000-2499999999999999999999
	2500000000000000000000-4999999999999999999999
	5000000000000000000000-9999999999999999999999
	10000000000000000000000-24999999999999999999999
	25000000000000000000000-49999999999999999999999
	50000000000000000000000-99999999999999999999999
	100000000000000000000000-249999999999999999999999
	250000000000000000000000-499999999999999999999999
	500000000000000000000000-999999999999999999999999
	1000000000000000000000000-2499999999999999999999999
	2500000000000000000000000-4999999999999999999999999
	5000000000000000000000000-9999999999999999999999999
	10000000000000000000000000-24999999999999999999999999
	25000000000000000000000000-49999999999999999999999999
	50000000000000000000000000-99999999999999999999999999
	100000000000000000000000000-249999999999999999999999999
	250000000000000000000000000-499999999999999999999999999
	500000000000000000000000000-999999999999999999999999999
	1000000000000000000000000000-2499999999999999999999999999
	2500000000000000000000000000-4999999999999999999999999999
	5000000000000000000000000000-9999999999999999999999999999
	10000000000000000000000000000-24999999999999999

Les unités primaires d'échantillonnage étaient les lieux de travail et un échantillon de $n=100$ de ces unités était soustrait. Cette taille d'échantillon était la seule acceptable étant donné le coût de l'enquête éventuelle. Les lieux de travail ont été stratifiés en fonction de deux variables, à savoir le type de branche d'activité (27 catégories) et le nombre d'employés (3 catégories). Les nombres prévus d'unités dans les cellules sous stratification proportionnelle sont donnés au tableau 7. Le plan réel d'échantillonnage utilisé pour l'étude était fondé sur la méthode de Bryant et coll. (1960) après avoir regroupé certaines strates, car il s'agissait de la seule méthode disponible au moment de la réalisation de l'étude.

Nous avons appliqué notre méthode à ce problème. La valeur minimale possible de $E[w(s)]$ en utilisant la méthode que nous proposons est 5,0418. Les étapes réelles de calcul sont les suivantes :

Première itération :

Etape 1. Tirer 500 échantillons pour former S^{n_0} , en générant au hasard les $n_i^{n_0}$, indépendamment pour chaque échantillon.

Etape 2. La valeur objective de (14) est 0,45088.

Deuxième itération :

Etape 1. Tirer 500 échantillons à ajouter à S^{n_0} .

Etape 2. La valeur objective de (14) est obtenue et produit la valeur minimale $E[w(s)] = 5,0418$.

Cette procédure a duré environ 30 secondes en se servant d'un programme Fortran sur poste de travail Sun Ultra 10.

Tableau 4
 Nombres prévus d'unités d'échantillonnage dans les
 cellules sous stratification proportionnelle ($n = 40$)

N° de ligne										N° de colonne																																																																																									
1										2										3										4										5										6										7										8										Marginal de ligne																			
1										2										3										4										5										6										7										8										Marginal de ligne																			
1										0,41										0,55										0,58										0,80										0,23										0,61										0,70										0,12										4									
2										0,52										0,15										0,07										0,90										0,28										0,10										0,37										0,61										3									
3										0,72										0,15										0,65										0,73										0,39										0,34										0,85										0,17										4									
4										0,70										0,55										0,46										0,10										0,41										0,05										0,24										0,49										3									
5										0,07										0,63										0,45										0,81										0,52										0,02										0,70										0,80										4									
6										0,61										0,33										0,79										0,21										0,02										0,61										0,67										0,76										4									
7										0,88										0,48										0,73										0,69										0,44										0,64										0,86										0,28										5									
8										0,22										0,14										0,85										0,37										0,69										0,45										0,49										0,79										4									
9										0,85										0,44										0,80										0,76										0,31										0,71										0,60										0,53										5									
10										0,02										0,58										0,62										0,63										0,71										0,47										0,52										0,45										4									
Total										10										10										10										10										10										10										10										10										40									
Marginal de										colonne										colonne										colonne										colonne										colonne										colonne										colonne										colonne										40									

Voici les étapes élémentaires de notre plan d'échantillonnage.

Étape 1. Obtenir un sous-ensemble d'échantillons candidats S_{n_0} par la méthode d'échantillonnage proposée (obtenus en trois minuscules). La proportion de l'échantillon peut être comparée à celle du tableau 4 pour déterminer dans quelle mesure elle s'approche de celle satisfaisant la propriété de RPP.

Étape 2. Résoudre le problème de programmation linéaire donné par (12) et (13) pour obtenir

$$\min_{\substack{p(s), s \in S_{n_0} \\ n_j}} \sum_{n_j} | \sum_{s \in S_{n_0}} n_j p(s) d(s) - n d_{ij} |. \quad (14)$$

Si la valeur objective de (14) est supérieure à zéro, répéter l'étape 1 avec un ensemble S_{n_0} plus grand. Si la valeur objective de (14) est nulle, arrêter, car une solution optimale a été obtenue.

N° de ligne										N° de colonne									
Marginal										Marginal									
1										2									
0,408										0,554									
0,582										0,776									
0,250										0,594									
0,734										0,102									
0,144										0,638									
0,692										0,542									
0,452										0,120									
0,416										0,044									
0,260										0,474									
0,708										0,016									
0,682										0,770									
0,480										0,734									
0,676										0,470									
0,664										0,842									
0,268										0,500									
0,486										0,502									
0,784										0,412									
0,692										0,624									
0,416										0,658									
0,564										0,026									
0,830										0,418									
0,870										0,026									
Total										Total									

Nous en donnons la preuve à l'annexe 1. Donc, si l'expression (14) devient nulle lorsqu'on applique la stratégie susmentionnée, la solution résultante donnera la valeur minimale de $E[w(s)]$ comme en (16).

Exemple 3. Exemple réel de matrice 27×3 avec valeurs de marge non entières : Nous illustrons la méthode au moyen d'un exemple réel tiré d'une étude de l'hygiène du milieu portant sur la santé au travail dans diverses branches d'activité en Finlande (Vihtma 1981). La population choisie pour l'étude comprenait 1 430 petits lieux de travail industriel (5 à 49 employés), représentant en tout 22 893 employés, dans la province d'Uusimaa, c'est-à-dire la province la plus au sud et la plus industrialisée de la Finlande.

$$E[w(s)] = \sum_{i=1}^I (r_i - \lfloor r_i \rfloor) (1 + \lfloor r_i \rfloor - r_i) + \sum_{j=1}^J (r_j - \lfloor r_j \rfloor) (1 + \lfloor r_j \rfloor - r_j). \quad (16)$$

Ceci, conjugué à la propriété de RPP, $E[n_j(s)] = \sum_{s \in S} n_j(s) p(s) = r_j$, implique que l'espérance de la fonction de manque de désirabilité $w(s)$ définie en (3) est constante pour $i = 1, \dots, R$, $j = 1, \dots, C$.

$$|n_i(s) - r_i| < 1 \quad \text{et} \quad |n_j(s) - r_j| < 1 \quad (15)$$

La méthode s'étend facilement aux valeurs de marge non entières. Il suffit simplement de remplacer n_j dans tout l'algorithme par n_j qui prend la valeur $\lfloor r_j \rfloor + 1$ avec une probabilité $\alpha = r_j - \lfloor r_j \rfloor$ et la valeur $\lfloor r_j \rfloor$ avec une probabilité $1 - \alpha$. La seule difficulté supplémentaire est que $E[w(s)]$ ne peut devenir nulle. Donc, nous n'avons pas de borne inférieure de référence évidente pour savoir si nous approchons ou non de la solution optimale. Cependant, la stratégie de randomisation susmentionnée garantit que, pour tout échantillon RPPA obtenu, nous ayons

3.4 Extension aux valeurs de marge non entières

Cette procédure a duré de 30 à 60 secondes en se servant d'un programme Fortran sur poste de travail Sun Ultra 10.

Étape 2. La valeur objective de (14) est nulle. Le plan d'échantillonnage final est obtenu.

Étape 1. Tirer 500 échantillons à ajouter à S_{n_0} .

Deuxième itération :

Étape 1. Tirer 500 échantillons pour former S_{n_0} .

Étape 2. La valeur objective de (14) est 0,1659.

Première itération :

Les étapes réelles de calcul sont les suivantes :

tableau 6.

Exemple 2. Matrice 20×15 avec valeurs de marge entières : Dans cet exemple, nous considérons la matrice

nulle.

Dans cet exemple, un sous-ensemble S_{n_0} candidat de 500 échantillons a suffi pour obtenir une valeur objective

Étape 2 : Poser $i = i + 1$.

Étape 2.1 : Pour $j = 1, \dots, C$, faire ce qui suit :

a) poser $R_j = \sum_{k=1}^K r_{kj}$

b) si $R_j - A_j \leq 0$ poser $a_{ij} = 0$,

c) si $R_j - A_j \geq 1$ poser $a_{ij} = 1$.

Étape 2.2 : Poser $J = \{j : 0 < R_j - A_j < 1\}$ et $not = \sum_{j=1}^J r_{ij} \times not / \sum_{j \in J} r_{ij}$, pour $j \in J$. S'il existe un $j_0 \in J$ tel

que $r_{ij_0} > 1$, alors poser $a_{ij_0} = 1$ et retourner à l'étape 2.1. Sinon, passer à l'étape 3.

Étape 3 : Tirer un échantillon de not cellules à partir de J par une méthode d'échantillonnage avec probabilités

inégales sans remise et r_{ij} pour obtenir a_{ij} pour $j \in J$.

Poser $A_j = \sum_{i=1}^I a_{ij}$ pour $j = 1, \dots, C$.

Étape 4 : Si $i = R$, alors s'arrêter; sinon, retourner à l'étape 2.

Notons qu'à l'étape 2, la façon de recalculer la i^e ligne de probabilités d'inclusion n'est pas unique. Cependant, les règles générales à suivre pour ce calcul sont :

a) $0 \leq r_{ij} \leq 1$ et, si $A_j = n_j$, autrement dit si suffisamment d'unités sont sélectionnées à partir de la j^e colonne, r_{ij} doit être fixé à 0; si $A_j = n_j - (R - i + 1)$, autrement dit si n'est pas possible de sélectionner un nombre suffisant d'unités à partir de cette colonne à moins que toutes les unités restantes soient sélectionnées, r_{ij} doit être fixé à 1;

b) garder $\sum_{j=1}^C r_{ij}' = \sum_{j=1}^C r_{ij} = n_i$.

La méthode s'étend facilement au cas des totaux marginaux non entiers. Nous reportons toutefois la discussion détaillée de cette extension plus loin. Nous pouvons maintenant utiliser la méthode susmentionnée pour produire un ensemble candidat, S_{n_0} , et lui appliquer la méthode de programmation linéaire. Pour comprendre pourquoi nous choisissons de modifier la méthode de programmation linéaire, il faut se rendre compte que, pour le cas des valeurs de marge entières, chaque $s \in S_{n_0}$ atteint déjà le minimum dans (2), de sorte qu'une application directe de la programmation linéaire revient à déterminer si l'existence ou non une solution possible. Donc, si nous gérons un S_{n_0} de taille égale à, disons, 500, puis 1 000, etc., et que le projeté de programmation linéaire continue de ne trouver aucune autre solution possible, nous ne savons vraiment pas si nous rapprochons ou non de la solution. Au lieu de cela, nous choisissons de renverser le problème d'optimisation et de résoudre un problème dual.

$$(8) \quad \min \sum_{i,j} | \sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} |.$$

Nous savons que $w(s) = 0$ pour tous $s \in S_{n_0}$ et nous cherchons une solution qui donne un nombre minimal de valeurs nulle dans (8). Essentiellement, nous avons permuté

les rôles de la fonction objective et de la contrainte de RPP moins aisé d'utiliser la programmation linéaire pour résoudre (8). Voici un moyen de le faire. Nous commençons par préciser les contraintes

$$\sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} + d_{ij} - e_{ij} = 0 \quad \text{pour } i = 1, \dots, R \quad \text{et } j = 1, \dots, C, \quad (9)$$

$$d_{ij} \geq 0, e_{ij} \geq 0, d_{ij}, e_{ij} = 0. \quad (10)$$

$$\text{Puis, nous notons que} \quad \left| \sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} \right| = \begin{cases} d_{ij} & \text{if } \sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} < 0 \\ e_{ij} & \text{if } \sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} \geq 0 \end{cases}$$

$$= d_{ij} + e_{ij} \quad (11)$$

Donc, nous pouvons remplacer (8) par

$$(12) \quad \min \sum_{p(s), d_{ij}, e_{ij}} (d_{ij} + e_{ij}),$$

$$\sum_{s \in S_{n_0}} n_{ij}(s) p(s) - r_{ij} + d_{ij} - e_{ij} = 0, d_{ij}, e_{ij}, p(s) \geq 0, d_{ij}, e_{ij} \quad (13)$$

3.3 Certains exemples pour des valeurs de marge entières

Nous allons maintenant utiliser deux exemples pour illustrer la méthode d'échantillonnage. Nous décrivons premier, c'est-à-dire une matrice 10×8 , en détail pour illustrer la procédure complète. Nous donnons le deuxième, pour une matrice de plus grande taille (20×15), pour illustrer la taille des problèmes qu'elle permet de résoudre (il s'agit presque de la taille limite des problèmes que la méthode proposée permet de résoudre en pratique). Toute méthode d'échantillonnage avec probabilités inégales sans remise peut y être intégrée. Dans l'exemple 1 qui suit, nous choisissons d'utiliser la méthode par groupement aléatoire de Rao, Hartley et Cochran (1962), puisqu'elle est simple et que nous ne devons vraiment reproduire qu'un proche-matériauement les probabilités de sélection, ce qu'elle permet de faire. Cependant, la méthode Rao-Hartley-Cochran ne donne de bons résultats que pour des problèmes 2 maximum, de taille moyenne. Dans le cas des exemples 2 et 3, nous devrions utiliser une méthode qui reproduit exactement les probabilités de sélection. Il en existe plusieurs, mais nous choisissons celle élaborée par Lu (2000).

Exemple 1. Matrice 10×8 avec valeurs de marge entières : Le tableau 4 donne un problème de stratification double avec le nombre prévu d'unités d'échantillonnage dans les cellules et la taille d'échantillon.

La méthode de programmation linéaire est simple et facile à appliquer. Son principal désavantage est d'ordre computationnel. Le nombre de paramètres du problème de programmation linéaire résultant est égal au nombre d'échantillons de taille n provenant des $RC > n$ cellules, $\binom{n}{RC}$, dont la grandeur devient assez rapidement impossible à traiter. À la section suivante, nous examinons divers moyens de rendre la méthode de programmation linéaire plus efficace sur le plan des calculs tout en retenant toutes ses bonnes propriétés.

3. MÉTHODE PRATIQUE DE PROGRAMMATION LINÉAIRE

L'idée fondamentale de la méthode de programmation linéaire consiste à obtenir un plan d'échantillonnage optimal en ce qui concerne le manque (minimal) prévu de « désirabilité » de l'échantillon en résolvant directement un problème de programmation linéaire dont les inconnues sont les $p(s), s \in S^n$, tout en maintenant la propriété de répartition proportionnelle prévue (RPP). Le seul obstacle est que le nombre d'éléments contenus dans S^n est souvent très grand et que, même avec la puissance de calcul moderne, il devient difficile d'effectuer la programmation linéaire si le nombre d'inconnues est grand.

Pour réduire la tâche de calcul pour ce problème de programmation linéaire déterminé par la cardinalité de S^n , nous voulons obtenir un sous-ensemble de S^n , disons $S_{n_0}^n$, qui soit presque aussi représentatif que S^n , mais beaucoup plus petit, et donc résoudre le problème de programmation linéaire qui suit en nous servant d'un ensemble beaucoup plus petit de $p(s), s \in S_{n_0}^n$ en tant qu'inconnues :

$$(6) \quad \min \sum_{s \in S_{n_0}^n} w(s) p(s).$$

3.1 Certaines stratégies justificatives

La stratégie susmentionnée est facile à énoncer, mais son application n'est pas entièrement évidente. En fait, nous pouvons explorer plusieurs directions pour déterminer un tel sous-ensemble $S_{n_0}^n \subset S^n$. Nous décrivons ici une méthode élémentaire liée aux fonctions de perte à laquelle ont fait allusion Sitter et Skinner (1994) et nous montrons comment elle permet d'augmenter modeste ment la taille des problèmes qui peuvent être résolus. Puis, nous discutons de certaines directions à prendre évidentes qui n'ont toutefois produit que peu d'amélioration. En décrivant ces efforts non fructueux, nous justifions la proposition éventuelle.

La principale souplesse de la méthode de programmation linéaire tient au choix de la fonction de perte $w(s)$. Donc, il est naturel que nous considérions d'abord la fonction de perte lorsque nous essayons

d'augmenter l'efficacité computationnelle de la méthode. L'examen de la fonction objective du problème de programmation linéaire (2) nous donne à penser qu'en tant que coefficients des inconnues $p(s)$, la fonction de perte $w(s)$ n'aura pas une valeur très grande lorsque la fonction objective sera minimisée. Autrement dit, dans un plan d'échantillonnage optimal, toutes les $p(s)$ positives seront attribuées uniquement à des échantillons dont le manque de « désirabilité » est faible. Partant de cette observation, nous émettons l'hypothèse qu'un bon remplacement pour S^n pourrait être le sous-ensemble qui suit

$$S_{n_0}^n = \{s \in S^n : w(s) = \sum_{i=1}^t (n_i^t(s) - n P_i^t)^2 - n P_i^t\}^2$$

$$(7) \quad + \sum_{j=1}^f (n_j^f(s) - n P_j^f)^2 - n P_j^f\}^2, \quad w_0\}$$

où w_0 est une constante positive prédéterminée. Dans le cas de valeurs de marge entières, on pourrait même poser $w_0 = 0$ et se limiter aux échantillons où les valeurs de marge concordent. Par exemple, au tableau 3, la solution n attribue une probabilité positive qu'à six échantillons et, pour chacun d'eux, la fonction objective est nulle.

Lu (2000) élabore des stratégies de programmation linéaire embodées pour résoudre ce problème. Pour des problèmes de taille modérée, comme des matrices 8×5 (c'est-à-dire, 40 cellules), cette méthode donne de bons résultats. Toutefois, pour des problèmes de plus grande amplitude, la taille des ensembles candidats résultants devient grande très rapidement, même dans le cas des valeurs de marge entières. Donc, pour les problèmes de grande taille, la méthode pose la même difficulté qu'auparavant, c'est-à-dire un ensemble candidat qui oblige à résoudre un problème de programmation linéaire où le nombre d'inconnues est trop élevé.

En réalité, même un ensemble d'échantillons candidats $S_{n_0}^n$ de la forme donnée par (7) est nettement plus grand qu'il n'est nécessaire pour que nous trouvions une solution optimale. Ce dont nous avons vraiment besoin est un sous-ensemble plus petit, mais suffisamment représentatif, où, par « petit », nous entendons suffisamment petit pour rendre possible la résolution du problème de programmation linéaire résultant et par « représentatif », nous entendons contenir les éléments qui promettent que ce problème de programmation linéaire soit faisable.

Avant de décrire la solution éventuelle que nous proposons pour ce problème, nous aimerions présenter certaines méthodes naïves d'obtention d'un tel « ensemble représentatif » qui n'ont pas donné de bons résultats. Ces méthodes ne sont pas très utiles en pratique, mais elles nous ont inspiré les idées qui sous-tendent la méthode plus élaborée que nous proposons.

1) Optimisation en deux étapes : Avant tout, dans (7), nous pourrions essayer de scinder $S_{n_0}^n$ en un grand nombre de sous-ensembles assez petits pour être traités individuellement par programmation linéaire, en espérant que les solutions optimales obtenues pour chacun à la première

biaisés des totaux des upe. Cependant, si nous devons sélectionner les upe avec probabilités inégales, disons $n z_{ijk}$ pour l'upe k dans la cellule de stratification ij , z_{ijk} sera habituellement égale à $M_{ijk} / \sum_{j,k} M_{ijk}$, où M_{ijk} représente une mesure de taille de l'upe k dans la cellule ij , nous pouvons facilement modifier la méthode en posant que P_{ij} est égale à z_{ij}/z_{\dots} , où $z_{ij} = \sum_k z_{ijk}$ et $z_{\dots} = \sum_{i,j,k} z_{ijk}$. Alors, si $n_{ij}(s) > 0$, nous sélectionnons un échantillon de $n_{ij}(s)$ upe dans la cellule ij selon une méthode donnée d'échantillonnage avec probabilité proportionnelle à z_{ijk} .

2.4 Exemple

Nous pouvons illustrer la méthode de programmation linéaire au moyen de l'exemple hypothétique de Bryant et coll. (1960) donné au tableau 1. Premièrement, nous simplifions le problème comme indiqué au tableau 2 afin de satisfaire l'hypothèse exprimée en (5). Puis, nous utilisons un logiciel type de programmation linéaire pour résoudre ce problème réduit (2). Puisque nous pouvons faire correspondre exactement les totaux marginaux entiers des nombres prévus d'unités d'échantillonnage dans les cellules aux totaux marginaux des tailles d'échantillon n_{ij} et $n_{\cdot j}$, ce qui signifie que la fonction de perte $w(s)$ peut prendre une valeur minimale nulle, la fonction objective en (2) pour cet exemple est, elle aussi, minimale lorsque sa valeur est nulle. La solution optimale de ce problème est donnée au tableau 3. Notons que cette solution a été reconstruite de façon à ce qu'elle concorde avec l'exemple original présenté au tableau 1.

Tableau 2

Exemple modifié de Bryant et coll. (1960)

Région	Type de collectivité	Type de collectivité				
		Urbanaine	Rurale	Métropolitaine	Total	
1		0,0	0,5	0,5	1,0	
2		0,2	0,3	0,5	1,0	
3		0,2	0,6	0,2	1,0	
4		0,6	0,8	0,6	2,0	
5		0,0	0,8	0,2	1,0	
Total		1,0	3,0	2,0	6,0	

Tableau 3

Solution par programmation linéaire de l'exemple tiré de Bryant et coll. (1960)

s	$d(s)$	s				
		1	2	3	4	5
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	1	1	0	1	0
5	0	0	1	1	0	1
6	0	0	0	1	1	0
7	0	0	0	0	1	1
8	0	0	0	0	0	1
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	0	0	0	0	0	0
26	0	0	0	0	0	0
27	0	0	0	0	0	0
28	0	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
31	0	0	0	0	0	0
32	0	0	0	0	0	0
33	0	0	0	0	0	0
34	0	0	0	0	0	0
35	0	0	0	0	0	0
36	0	0	0	0	0	0
37	0	0	0	0	0	0
38	0	0	0	0	0	0
39	0	0	0	0	0	0
40	0	0	0	0	0	0
41	0	0	0	0	0	0
42	0	0	0	0	0	0
43	0	0	0	0	0	0
44	0	0	0	0	0	0
45	0	0	0	0	0	0
46	0	0	0	0	0	0
47	0	0	0	0	0	0
48	0	0	0	0	0	0
49	0	0	0	0	0	0
50	0	0	0	0	0	0
51	0	0	0	0	0	0
52	0	0	0	0	0	0
53	0	0	0	0	0	0
54	0	0	0	0	0	0
55	0	0	0	0	0	0
56	0	0	0	0	0	0
57	0	0	0	0	0	0
58	0	0	0	0	0	0
59	0	0	0	0	0	0
60	0	0	0	0	0	0
61	0	0	0	0	0	0
62	0	0	0	0	0	0
63	0	0	0	0	0	0
64	0	0	0	0	0	0
65	0	0	0	0	0	0
66	0	0	0	0	0	0
67	0	0	0	0	0	0
68	0	0	0	0	0	0
69	0	0	0	0	0	0
70	0	0	0	0	0	0
71	0	0	0	0	0	0
72	0	0	0	0	0	0
73	0	0	0	0	0	0
74	0	0	0	0	0	0
75	0	0	0	0	0	0
76	0	0	0	0	0	0
77	0	0	0	0	0	0
78	0	0	0	0	0	0
79	0	0	0	0	0	0
80	0	0	0	0	0	0
81	0	0	0	0	0	0
82	0	0	0	0	0	0
83	0	0	0	0	0	0
84	0	0	0	0	0	0
85	0	0	0	0	0	0
86	0	0	0	0	0	0
87	0	0	0	0	0	0
88	0	0	0	0	0	0
89	0	0	0	0	0	0
90	0	0	0	0	0	0
91	0	0	0	0	0	0
92	0	0	0	0	0	0
93	0	0	0	0	0	0
94	0	0	0	0	0	0
95	0	0	0	0	0	0
96	0	0	0	0	0	0
97	0	0	0	0	0	0
98	0	0	0	0	0	0
99	0	0	0	0	0	0
100	0	0	0	0	0	0

l'erreur quadratique moyenne de l'estimateur \bar{y} sous un modèle d'analyse de la variance (voir Sitter et Skinner, 1994). Alors, en résolvant le problème de programmation linéaire susmentionné, nous pouvons obtenir une EQM minimisée au sens de l'analyse de la variance tout en maintenant la propriété de répartition proportionnelle prévue (RPP) de $n_{ij}(s)$. Notons que, si nous obtenons un plan d'échantillonnage dont la fonction objective est nulle, toutes les contraintes de marge sont respectées. En général, cette situation ne se présente que dans le cas où les totaux marginaux sont des nombres entiers.

Selon Sitter et Skinner (1994), un moyen simple de réduire la taille de S_n consiste à limiter les valeurs réelles que peut prendre n_{ij} à $\lfloor nP_{ij} \rfloor$ ou $\lfloor nP_{ij} \rfloor + 1$, où $\lfloor nP_{ij} \rfloor$ est le plus grand nombre entier inférieur ou égal à nP_{ij} . En écrivant $n_{ij} = n_{ij} - \lfloor nP_{ij} \rfloor$ et $r_{ij} = nP_{ij} - \lfloor nP_{ij} \rfloor$, nous pouvons alors imposer

$$E(n_{ij}) = r_{ij} \quad (4)$$

où $n_{ij} = 0$ ou 1 et $0 < r_{ij} < 1$. Alors, nous pouvons appliquer la méthode de programmation linéaire aux n_{ij} et, finalement, nous pouvons utiliser $\lfloor nP_{ij} \rfloor + n_{ij}$ comme taille réelle des échantillons de cellule. Par conséquent, sans perte de généralité, nous supposons que

$$n_{ij} = 0, 1 \text{ et } 0 < r_{ij} = nP_{ij} < 1. \quad (5)$$

2.2 Stratification plus profonde

La méthode de Sitter et Skinner (1994) s'étend facilement à un plus grand nombre de variables de stratification si l'on pose que s représente la matrice correspondant à r dimensions. La fonction de perte inclura alors un plus grand nombre de termes; par exemple, pour une stratification triple, l'équation (3) pourrait être remplacée par

$$w(s) = \gamma_1 \sum_{R_1}^t (n_{\cdot \cdot \cdot}(s) - nP_{\cdot \cdot \cdot})^2 + \gamma_2 \sum_{R_2}^f (n_{\cdot j \cdot}(s) - nP_{\cdot j \cdot})^2 + \gamma_3 \sum_{R_3}^k (n_{i \cdot \cdot}(s) - nP_{i \cdot \cdot})^2$$

2.3 Échantillonnage à plusieurs degrés

Une application importante de la stratification multiple est la sélection d'unités primaires d'échantillonnage (upe) en cas d'échantillonnage à plusieurs degrés, où il est courant de disposer de plusieurs variables de stratification. À la section 2.1, la probabilité d'inclusion de chaque upe avec probabilités égales, nous pouvons étendre directement la méthode en remplaçant les unités par les valeurs observées de y , par des estimateurs non

simple, est adaptable à diverses situations, produit toujours une solution et donne une EQM dont les propriétés sont meilleures. Sa principale limite pratique est que les calculs deviennent ardues à mesure qu'augmente le nombre de cellules de la stratification multiple, et sont rapidement infaisables. Dans le présent article, nous représentons une méthode simple qui permet d'appliquer la méthode de programmation linéaire à la résolution de problèmes de taille nettement plus grande. À la section 2, nous décrivons la méthode de programmation linéaire de Sitter et Skinner (1994) pour introduire la notation et nous discutons brièvement de ses limites numériques. À la section 3.1, nous examinons d'abord certaines stratégies simples qui permettent de rendre les calculs moins ardues à titre de justification de la proposition éventuelle. Aux sections 3.2 et 3.3, nous discutons de la méthode proposée, en supposant que les totaux marginaux sont des nombres entiers et nous donnons certains exemples où le nombre de cellules de stratification varie de 80 à 300 pour illustrer la capacité qu'à la nouvelle méthode de traiter des problèmes de grande taille. À la section 3.4, nous décrivons l'extension simple de la méthode aux totaux marginaux non entiers et nous l'illustrons en l'appliquant à un exemple réel tiré de données publiées sur la santé au travail (Vijhima 1981).

Tableau 1
Exemple tiré de Bryant et coll. (1960). Nombres prévus d'unités d'échantillonnage dans les cellules sous stratification proportionnelle ($n = 10$)

Région	Type de collectivité				
	Total	Urbaine	Rurale	Métropolitaine	Total
1	1,0	0,5	0,5	2,0	2,0
2	0,2	0,3	0,5	1,0	1,0
3	0,2	0,6	1,2	2,0	2,0
4	0,6	1,8	0,6	3,0	3,0
5	1,0	0,8	0,2	2,0	2,0
Total	3,0	4,0	3,0	10,0	10,0

2. LA MÉTHODE DE PROGRAMMATION LINÉAIRE

2.1 Notions fondamentales

Nous considérons, pour introduire la méthode de programmation linéaire de Sitter et Skinner (1994), la forme la plus simple de stratification double. Supposons que N unités d'une population finie sont disposées, conformément à une classification à double entrée, en R lignes correspondant aux catégories de l'autre variable, et en C colonnes correspondant aux catégories de l'autre variable. Représentons par N_{ij} le nombre d'unités de la population dans la i^{e} ligne et la j^{e} colonne (c'est-à-dire, dans la ij^{e} cellule) du tableau à double entrée et par $P_{ij} = N_{ij}/N$, la proportion de la population totale comprise dans la ij^{e} cellule. Soit X , la valeur moyenne d'une caractéristique étudiée y pour la population et X_{ij} la valeur moyenne de y pour la ij^{e} cellule.

Nous sélectionnons l'échantillon comme suit.

i) Nous déterminons la taille d'échantillon n_{ij} aléatoirement pour chaque cellule, conformément à une méthode précisée a priori. Si nous représentons par s la matrice $R \times C$ (n_{ij} , $i = 1, \dots, R$, $j = 1, \dots, C$), nous attribuons par cette méthode une probabilité $p(s)$ à chaque s compris dans l'ensemble S de telles matrices possibles et nous sélectionnons une matrice unique, s , à partir de S . Pour représenter la dépendance de n_{ij} à l'égard de s , nous écrivons $n_{ij}(s)$.

ii) Puis, nous sélectionnons un échantillon aléatoire simple de $n_{ij}(s)$ unités dans la ij^{e} cellule et nous obtenons les valeurs de y .

Nous nous limitons aux plans d'échantillonnage à taille d'échantillon constante $n > 0$, autrement dit, à des matrices $s \in S_n$ telles que $\sum_{i=1}^R \sum_{j=1}^C n_{ij}(s) = n$. Nous aimerions aussi nous limiter à la stratification proportionnelle de sorte que

$$\sum_{s \in S_n} n_{ij}(s) p(s) = n P_{ij} \quad \text{pour } i = 1, \dots, R, j = 1, \dots, C, (1)$$

qui implique que la moyenne d'échantillon simple non pondérée $\bar{y}(s)$ est un estimateur non biaisé de X . Nous donnons à (1) le nom de contrainte de répartition proportionnelle prévue (RPP).

La méthode de programmation linéaire de Sitter et Skinner (1994) consiste à choisir un plan d'échantillonnage $p(s)$ qui minimise le manque prévu de « désirabilité » des échantillons par résolution du problème de programmation linéaire :

$$\min \sum_{s \in S_n} w(s) p(s) \quad (2)$$

où $w(s)$ est une fonction de subordonné à la contrainte (1), où $w(s)$ est une fonction de perte pour l'échantillon s , qu'il convient de spécifier, et où les probabilités $p(s)$ sont inconnues. Sitter et Skinner (1994) ont exploité l'observation principale de Rao et Nigam (1990, 1992) lors de travaux visant à éviter les échantillons indésirables, à savoir que, dans (2), la fonction objective est linéaire en probabilités $p(s)$ (voir aussi Lahiri et Mukerjee, 2000).

Dans la fonction objective (2), la fonction de perte $w(s)$ joue un rôle important. Si $w(s)$ est bien définie, nous avons la souplesse de rechercher l'existence d'une solution optimale de (2) dans un ensemble S_n de taille économique, et, par-dessus tout, de rendre l'estimation plus efficace. Sitter et Skinner (1994) proposent de choisir

$$w(s) = \sum_{i=1}^R \sum_{j=1}^C \left(n_{ij}(s) - n P_{ij} \right)^2 + \sum_{j=1}^C \left(n_{.j}(s) - n P_{.j} \right)^2, \quad (3)$$

où $n_{.j}(s) = \sum_{i=1}^R n_{ij}(s)$, $n_{ij}(s) = \sum_{i=1}^R n_{ij}(s)$, $P_{.j} = \sum_{i=1}^R P_{ij}$, et $P_{ij} = \sum_{i=1}^R P_{ij}$. De toute évidence, la fonction objective (2) est effectivement $E(w(s))$ pour toute $p(s)$ déterminée par le plan de sondage et peut être interprétée comme étant

Méthode pratique de stratification multiple par programmation linéaire

WILSON LU et RANDY R. SITTER¹

RÉSUMÉ

Sitter et Skinner (1994) présentent une méthode qui consiste à appliquer la programmation linéaire à la conception d'enquêtes avec stratification multiple, principalement dans des situations où la taille souhaitée de l'échantillon est inférieure ou à peine supérieure au nombre total de cellules de stratification. Leur méthode repose sur une idée simple, facile à comprendre et à appliquer. Cependant, en pratique, elle a le désavantage de devenir rapidement coûteuse en raison de l'importance des calculs, à mesure qu'augmente le nombre de cellules de la stratification multiple, au point de ne pouvoir être utilisée dans la plupart des situations réelles. Dans le présent article, nous étendons cette méthode de programmation linéaire et élaborons des méthodes en vue de réduire le nombre de calculs, de sorte qu'il soit possible de résoudre des problèmes de grande taille.

MOTS CLÉS : Échantillonnage PPT; répartition proportionnelle; groupement aléatoire; échantillonnage.

1. INTRODUCTION

En pratique, dans le cas de nombreuses enquêtes, il existe plusieurs variables de stratification et le concepteur de l'enquête peut donc choisir de définir des strates sous forme de cellules obtenues par recoupement des catégories de ces variables. À cet égard, consulter, par exemple, Engle, Marden et Pollock (1971), Hess, Riedel et Fitzpatrick (1976), Vihma (1981) et Skinner, Holmes et Holt (1994). Cette stratification multiple donne souvent lieu à des situations où la taille souhaitée de l'échantillon est inférieure ou à peine supérieure au nombre total de cellules de stratification (phénomène particulièrement courant lors du choix des unités primaires d'échantillonnage (upe) dans le cas de plans de sondage stratifiés à plusieurs degrés), ce qui empêche parfois d'appliquer les méthodes classiques de répartition de l'échantillon entre les strates.

Nous présentons au tableau 1 une illustration fondée sur un exemple hypothétique de Bryant, Hartley et Jessen (1960). Les collectivités (upe) sont classées en fonction de deux variables de stratification, le type et la région, comportant trois et cinq catégories, respectivement. La taille souhaitée de l'échantillon, $n = 10$, est inférieure au nombre total de cellules, c'est-à-dire 15. Cet exemple illustre aussi un problème connexe. Dans le tableau 1, les entrées sont les dénombrements prévus en cas de stratification proportionnelle, c'est-à-dire quand les tailles d'échantillon de strate sont proportionnelles aux tailles de population de strate. En général, à cause des contraintes de taille d'échantillon, les nombres prévus d'unités d'échantillonnage dans les cellules ne sont pas entiers. Si l'unité près ne produira pas de bons choix et entraînera sérieusement la contrainte de répartition proportionnelle. Il

Cette méthode de programmation linéaire est de conception

trier parti de la puissance des méthodes modernes de calcul.

propose une méthode de programmation linéaire qui vise à

aussi Lahiri et Mukerjee, 2000), Sitter et Skinner (1994) ont

(1990, 1992) afin d'éviter des échantillons indésirables (voir

solution. En s'inspirant de l'idée exposée par Rao et Nigam

Causey, Cox et Ernst (1985), pourraient ne pas donner de

assez difficiles à appliquer et, comme le font remarquer

méthodes de stratification double et triple, mais elles sont

treillis », par exemple Jessen 1973, 1975). Il propose deux

les méthodes connexes à la rubrique « échantillonnage en

cela pourrait être souhaitable dans certaines situations (voir

forcer la taille de cellules spécifiques à être nulle, alors que

Bryant et coll. (1960) tient à ce qu'il est impossible de

Selon Jessen (1970), une autre limite de la méthode de

pas être satisfaisantes (consulter Sitter et Skinner, 1994).

les propriétés de l'EQM de ces estimateurs pourraient ne

probabilités des distributions marginales sont respectées),

sélection contrôlée convenable, puisque seules les

cellule (autrement dit, la méthode n'est pas une méthode de

comme de la représentation proportionnelle de chaque

tailles prévues d'échantillon de cellule ne tenaient pas

sur ce scénario d'échantillonnage. Toutefois, puisque les

stratification double et ont donné deux estimateurs fondés

de la taille d'échantillon à chaque cellule en cas de

ont présenté une méthode fort simple d'attribution aléatoire

Fitzpatrick, 1961; Waterton, 1983). Bryant et coll. (1960)

échantillonnage systématiques aléatoires (voir Hess, Riedel et

appelée sélection contrôlée, par une méthode d'échan-

résoudre ce problème selon une démarche qu'ils ont

Goodman et Kish (1950) ont été les premiers à essayer de

pas entiers, ce qui peut aussi causer des difficultés.

est également courant que les totaux marginaux ne soient

¹ Wilson Lu, étudiant de doctorat, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; Randy R. Sitter, professeur, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

HIDIROGLOU, M.A., et SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

JOHNSON, N.T., et KOTZ, S. (1970). *Continuous Univariate Distribution-I*. New York: John Wiley & Sons, Inc.

LAVALLÉE, P., et HIDIROGLOU, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.

OSLO, I.T. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*, 23, 15-25.

RIVEST, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-72.

SÄRNDAHL, C.-E., SWENSSON, B., et WRETMAN, I. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

SCHNEEBERGER, H. (1979). Saddle points of the variance of the sample mean in stratified sampling. *Sankhyā : The Indian Journal of Statistics, Series C*, 41, 92-96.

SERFLING, R.J. (1968). Approximate optimal stratification. *Journal of the American Statistical Association*, 63, 1298-1309.

SETHI, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

SINGH, R.J. (1971). Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*, 66, 829-834.

SINGH, R., et PARKASH, D. (1975). Optimal stratification for equal allocation. *Annals of the Institute of Statistical Mathematics*, 27, 273-280.

SINGH, R., et SUKATME, B.V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*, 21, 515-528.

SLANTA, J., et KRENZKE, T. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 693-698.

SLANTA, J., et KRENZKE, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau du Census. *Techniques d'enquête*, 22, 65-75.

WANG, M.C., et AGGARWAL, V. (1984). Stratification under a particular Pareto distribution. *Communications in Statistics, Part A - Theory and Methods*, 13, 711-735.

YAVADA, S., et SINGH, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics, Part A - Theory and Methods*, 13, 2793-2806.

REMERCIEMENTS

L'auteur remercie Nathalie Vandal et Gaëtan Dalgé qui ont programmé les fonctions SAS IML pour les algorithmes de stratification utilisés dans le présent article. Il remercie aussi le rédacteur en chef et l'examineur de leurs commentaires constructifs.

BIBLIOGRAPHIE

ANDERSON, D.W., KISH, L. et CORNELL, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

COCHRAN, W.G. (1977). *Sampling Techniques*. Troisième édition. New York : John Wiley & Sons, Inc.

DALÉNUS, T. (1952). The Problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*, 35, 61-70.

DALÉNUS, T., et GURNÉY, M. (1951). The Problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, 34, 133-148.

DORMAN, A.H., et VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.

ECKMAN, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.

GOREY, J., ROSHWALB, A. et WRIGHT, R.L. (1984). Model-based stratification in inventory cost estimation. *Journal of Business and Economic Statistics*, 2, 1-9.

HEDLIN, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

HIDIROGLOU, M. (1994). Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 153-162.

Les dérivées partielles nécessaires pour appliquer l'algorithme itératif de Sethi (1963) sont

$$\begin{aligned}\frac{\partial W_h}{\partial n} &= \frac{A e^{\psi_h} / (e^{\psi_h} W_h - \phi_h^2)^{1/2}}{A^2 (\phi_h / W_h)^2 / N} - \frac{F}{F^2} \\ \frac{\partial \phi_h}{\partial n} &= \frac{-2A \phi_h / (e^{\psi_h} W_h - \phi_h^2)^{1/2}}{2A^2 \phi_h / (W_h N)} + \frac{F}{F^2} \\ \frac{\partial n}{\partial \phi_h} &= \frac{e^{\psi_h} A W_h / \{e^{\psi_h} W_h - \phi_h^2\}^{1/2}}{e^{\sigma^2} A^2 / N} - e^{\sigma^2} \frac{F}{F^2},\end{aligned}$$

et

$$F = \left(\sum x_i^2 / N \right)^2 c^2 + \sum_{i=1}^h c^2 \psi_h - \phi_h^2 W_h^2 / N.$$

6. CONSIDÉRATION NUMÉRIQUES

Slania et Krenzke (1994, 1996) ont éprouvé des difficultés d'ordre numérique lors de l'utilisation de l'algorithme de Lavallée et Hidiroglou avec la répartition optimum de Neyman : la convergence était lente et l'algorithme ne convergait pas toujours vers la valeur minimale réelle de n . En effet, Schneebberger (1979) et Slania et Krenzke (1994) ont montré que, pour une population bino-

miale particulière, le problème présente un col; autrement dit, les dérivées partielles sont toutes nulles aux limites b_h qui ne donnent pas de valeur minimale réelle de n . Lors de l'utilisation des algorithmes présentés ici, nous avons aussi éprouvé les difficultés numériques évoquées dans Slania et Krenzke (1994). Ceux construits sous la répartition puissance sont généralement plus stables que ceux construits sous la répartition optimum de Neyman; les difficultés numériques sont fréquentes lorsque le nombre L de strates est grand. En outre, à mesure que la distribution de X s'écarte de celle de X , c'est-à-dire à mesure que σ^2 augmente, la non-convergence de l'algorithme et l'impossibilité d'atteindre la valeur minimale globale de n deviennent plus fréquentes. Dans ces situations, les valeurs de départ de l'algorithme de stratification jouent un rôle de la plus grande importance. Par exemple, dans le tableau 2, le plan X obtenu dans le cas de la répartition optimum de Neyman dépend fortement des valeurs de départ. L'algorithme présenté au tableau 2 a comme valeurs de départ les limites choisissant pour valeurs de départ de l'algorithme les

limites obtenues au tableau 1 pour l'algorithme de Lavallée et Hidiroglou avec répartition optimum de Neyman, nous obtenons un plan d'échantillonnage différent pour lequel $n = 29$. Une bonne stratégie de calcul consiste à exécuter l'algorithme de stratification pour plusieurs plans d'échantillonnage intermédiaires afin d'obtenir un plan d'échantillonnage final, en utilisant les limites de strate obtenues lors d'une étape comme valeurs de départ de l'algorithme à l'étape suivante. L'application de l'algorithme log-linéaire se fait toujours en deux étapes. On commence par exécuter l'algorithme de Lavallée et Hidiroglou en fixant $\sigma = 0$, puis on utilise ces limites comme valeurs de départ pour l'exécution de la variable étudiée. Deux modèles statistiques sont introduits à cette fin. La nouvelle classe d'algorithmes s'appuie sur la règle de dérivation d'une fonction composée, ou règle d'enchaînement, pour calculer les dérivées partielles et sur la méthode de Sethi (1963) pour déterminer les limites optimales de strate.

7. CONCLUSION

Le présent article propose des généralisations de l'algorithme de stratification à modèle log-linéaire proposé dans l'article a été utilisé avec de bons résultats dans plusieurs enquêtes conçues par le Service de consultation statistique de l'Université Laval. Pour estimer la production annuelle totale de stop d'étrable, le nombre variable de taille pratique. Nous avons utilisé des données historiques pour estimer les paramètres du modèle log-linéaire reliant les étables producteurs de sève et le volume produit. Un autre exemple est l'estimation du déficit total au titre de l'entretien des bâtiments hospitaliers au Québec. La valeur de chaque immeuble a été choisie comme variable de stratification connue. Des experts ont estimé le déficit relatif à l'entretien des bâtiments comme étant de l'ordre de 20 % à 40 %. La résolution de $4\sigma^{\log} = \log(40\%) - \log(20\%)$ donne $\sigma^{\log} = \log(2)/4 = 0,17$ comme valeur paramétrique possible pour le modèle log-linéaire de la section 3.1. Dans ces deux exemples, le fait de tenir compte des divergences entre la variable de stratification et la variable étudiée augmente la taille de l'échantillon n d'un c.v. estimés sont proches des c.v. cibles.

Deux fonctions SAS IML appliquant l'algorithme décrit dans le présent article, l'une pour la répartition par la

de la fonction quadratique correspondante ayant la valeur la plus grande. Si $h < L - 1$, ceci nous donne

$$= \frac{b_{\beta \text{ nouv}}^h}{b_{\beta \text{ nouv}}^{h+1}} \left(- \frac{\partial \phi_h}{\partial n} - \frac{\partial \phi_{h+1}}{\partial n} \right) / \left(2 \left(\frac{\partial \psi_h}{\partial n} - \frac{\partial \psi_{h+1}}{\partial n} \right) \right)$$

$$+ \left(\frac{\partial \phi_h}{\partial n} - \frac{\partial \phi_{h+1}}{\partial n} \right)^2 - 4 \left(\frac{\partial \psi_h}{\partial n} - \frac{\partial \psi_{h+1}}{\partial n} \right) \left(\frac{\partial \phi_h}{\partial n} - \frac{\partial \phi_{h+1}}{\partial n} \right) \left(\frac{\partial \psi_h}{\partial n} - \frac{\partial \psi_{h+1}}{\partial n} \right)$$

tandis que si $h = L - 1$, nous avons

$$\frac{b_{\beta \text{ nouv}}^{L-1}}{b_{\beta \text{ nouv}}^L} = \frac{\left(\frac{\partial \phi_{L-1}}{\partial n} - \frac{\partial \phi_L}{\partial n} \right)^2 - 4 \left(\frac{\partial \psi_{L-1}}{\partial n} - \frac{\partial \psi_L}{\partial n} \right) \left(\frac{\partial \phi_{L-1}}{\partial n} - \frac{\partial \phi_L}{\partial n} \right) \left(\frac{\partial \psi_{L-1}}{\partial n} - \frac{\partial \psi_L}{\partial n} \right)}{\left(\frac{\partial \phi_{L-1}}{\partial n} - \frac{\partial \phi_L}{\partial n} \right)^2 - 4 \left(\frac{\partial \psi_{L-1}}{\partial n} - \frac{\partial \psi_L}{\partial n} \right) \left(\frac{\partial \phi_{L-1}}{\partial n} - \frac{\partial \phi_L}{\partial n} \right) \left(\frac{\partial \psi_{L-1}}{\partial n} - \frac{\partial \psi_L}{\partial n} \right)}$$

Les dérivées partielles de n par rapport à $W_h \phi_h$, et ψ_h dépendent des moments d'ordre 0, 1 et 2 de x_{β} dans la strate h . Nous calculons ces moments d'après les N valeurs de x dans la population. Par exemple,

$$\phi_h = \frac{1}{N} \sum_{i: b_{h-1} < x_i \leq b_h} x_i^{\beta}.$$

Nous donnons plus loin des applications de cette méthode

générale.

Lors de l'utilisation de l'algorithme de Sethi, on pose

habituellement que $L \geq 3$. Notons toutefois que l'algorithme marche aussi si $L = 2$. Dans ce cas, l'algorithme recherche la limite entre une strate à tirage complet et une strate à tirage partiel. Les évaluations successives de $b_{\beta \text{ nouv}}^{L-1}$ présentes plus haut produisent une limite optimale. Quand on suppose que la variable de stratification et la variable étudiée coïncident, c'est-à-dire $X = Y$, cette limite est presque identique à celle obtenue au moyen de l'algorithme présenté dans Hidiroglou (1986).

5.2 Un algorithme pour la répartition par la méthode puissance

Pour le modèle log-linéaire de la section 3.1, l'espérance conditionnelle est $E(X|b_h \geq X > b_{h-1}) = C \phi_h / W_h$ tandis que la variance conditionnelle est

$$\text{Var}(X|b_h \geq X > b_{h-1}) = C^2 \{ e^{\sigma^2 \psi_h} / W_h - (\phi_h / W_h)^2 \},$$

où $C = \exp(\alpha + \sigma^2/2)$. Aux termes de la règle de répartition par la méthode puissance, $a_h x = \phi_h^h / \sum_{h=1}^{L-1} \phi_h^h$, et la formule (5.7) pour n devient

$$n = N W^L + \frac{\left(\sum_{h=1}^{L-1} x_{\beta}^i / N \right)^2 c^2 + \sum_{h=1}^{L-1} e^{\sigma^2 \psi_h} W_h - \phi_h^2 / W_h}{\left(\sum_{h=1}^{L-1} e^{\sigma^2 \psi_h} W_h - \phi_h^2 \right)^{1/2}}.$$

et la formule pour n est

$$a_h x = \frac{\left\{ e^{\sigma^2 \psi_h} W_h - \phi_h^2 \right\}^{1/2}}{\sum_{h=1}^{L-1} \left\{ e^{\sigma^2 \psi_h} W_h - \phi_h^2 \right\}^{1/2}}.$$

Dans le cas de la répartition optimum de Neyman, la règle de répartition (2.3) écrite en fonction de $W_h \phi_h$, et ψ_h est

5.3 Un algorithme pour la répartition optimum de Neyman

$$F = \left(\sum_{h=1}^{L-1} x_{\beta}^i / N \right)^2 c^2 + \sum_{h=1}^{L-1} e^{\sigma^2 \psi_h} W_h - \phi_h^2 / W_h / N.$$

et

$$A = \sum_{h=1}^{L-1} \phi_h^h, B = \sum_{h=1}^{L-1} \left(e^{\sigma^2 \psi_h} W_h - \phi_h^2 \right) / \phi_h^h,$$

où

$$\frac{\partial \phi_h}{\partial n} = e^{\sigma^2 A W_h / \phi_h^h} \frac{F}{F^2} - e^{\sigma^2 A B / N},$$

$$+ \frac{F^2}{A B \phi_h^h / (n W_h)}$$

$$\frac{\partial \phi_h}{\partial n} = \frac{F}{A \{ -p e (\sigma^2 W_h \psi_h - \phi_h^2) / \phi_h^{h+1} - 2 / \phi_h^{h-1} \} + p \phi_h^{h-1} B}.$$

$$\frac{\partial W_h}{\partial n} = \frac{F}{A e^{\sigma^2 \psi_h} / \phi_h^h} - \frac{F}{A B (\phi_h / W_h)^2 / N}.$$

$h \leq L - 1$.

Les dérivées partielles nécessaires pour appliquer l'algorithme de stratification se calculent facilement; pour

$$n = N W^L + \frac{\left(\sum_{h=1}^{L-1} x_{\beta}^i / N \right)^2 c^2 + \sum_{h=1}^{L-1} e^{\sigma^2 \psi_h} W_h - \phi_h^2 / W_h}{\sum_{h=1}^{L-1} \left(e^{\sigma^2 \psi_h} W_h - \phi_h^2 \right) / \phi_h^h}.$$

Nous avons également appliqué à la variable REV84 l'algorithme de stratification choisi pour le modèle à remplacement aléatoire de la section 3.3 (avec répartition optimum de Neyman). Si l'on suppose qu'il existe des variations pour 2 % des unités ($\varepsilon = 0,02$), l'algorithme généralisé de Lavallée et Hidiroglou produit un plan d'échantillonnage stratifié avec $n = 37$ unités d'échantillonnage, l'estimateur résultant du total de RMT85 à un c.v. de 5,5 %. Une propriété intéressante de ce plan d'échantillonnage stratifié est que la fraction d'échantillonnage la plus faible est $\min h_i f_i = 9,3$ %; cette valeur est nettement plus grande que celle de $\min h_i f_i$ pour les plans d'échantillonnage des tableaux 1 et 2. Malgré l'existence de valeurs extrêmes, le modèle à remplacement aléatoire ne décrit pas aussi bien que le modèle log-linéaire les divergences entre REV84 et RMT85. C'est pourquoi une plus grande taille d'échantillon, 37 au lieu de 28, est nécessaire pour obtenir un estimateur dont la variance est comparable à celle obtenue pour la stratification basée sur un modèle log-linéaire.

5. UNE MÉTHODE DE CONSTRUCTION D'ALGORITHMES DE STRATIFICATION

Le but d'un algorithme de stratification est de déterminer les limites de strate et les tailles d'échantillon optimales pour l'échantillonnage de Y en se servant des valeurs communes $\{x_i, i = 1, \dots, N\}$ de la variable X pour toutes les unités de la population. Un modèle, comme ceux décrits à la section 3, caractérise la relation entre X et Y . Dans cette section, nous étendons l'algorithme de stratification de Lavallée et Hidiroglou (1988) à des situations où X et Y diffèrent. Nous servons du modèle log-linéaire de la section 3.1 pour tenir compte des différences entre Y et X . Les modifications nécessaires pour traiter le modèle à remplacement aléatoire sont faciles à appliquer (voir Rivest

5.1 Une généralisation de la méthode de stratification de Sethi (1963)

Il est pratique de considérer une population infinie analogue à l'équation (2.4) pour n . Puisque la variable aléatoire X a une densité $f(x)$, les deux premiers moments conditionnels de Y , étant donné que $b_{h-1} < X \leq b_h$, peuvent s'écrire en fonction de

$$W_h = \int_{b_{h-1}}^{b_h} f(x) dx, \quad \Phi_h = \int_{b_{h-1}}^{b_h} x f(x) dx, \quad \text{et } \psi_h = \int_{b_{h-1}}^{b_h} x^2 f(x) dx,$$

où β est la pente du modèle log-linéaire donné à la section 3.1 (à la présente section, β et σ représentent les paramètres du modèle log-linéaire de la section 3.1, mais, puisqu'il n'y a aucun risque de confusion, nous n'utilisons

plus l'indice \log). Aux fins de la stratification, il est utile de réécrire (2.4) en fonction des moyenne et variance conditionnelles de Y ,

$$n = NW_L + \frac{\sum_{L-1}^L W_h^2 \text{Var}(Y|b_h \leq X < b_{h-1})/a_h X}{\sum_{L-1}^L W_h^2 \text{Var}(Y|b_h \leq X < b_{h-1})/N}, \quad (5.7)$$

où $a_h X$ représente la règle de répartition écrite en fonction de la variable connue X . Par exemple, dans le cas de la répartition par la méthode puissance,

$$a_h X = \frac{\{W_h^k E(Y|b_h \leq X < b_{h-1})\}^p}{\sum_{L-1}^k \{W_h^k E(Y|b_h \leq X < b_{h-1})\}^p},$$

pour $h = 1, \dots, L-1$. Étant donné un modèle de la relation entre Y et X , $\text{Var}(Y|b_h \leq X < b_{h-1})$ et $E(Y|b_h \leq X < b_{h-1})$ peuvent s'écrire en fonction de W_h^h, Φ_h , et ψ_h . Donc, nous pouvons évaluer les dérivées partielles de n par rapport à b_h , pour $h < L-1$, en appliquant la règle de dérivation d'une fonction composée, ou règle d'enchaînement,

$$\frac{\partial n}{\partial W_h} = \frac{\partial W_h}{\partial W_h} + \frac{\partial \Phi_h}{\partial W_h} + \frac{\partial \psi_h}{\partial W_h}, \quad \frac{\partial n}{\partial \Phi_h} = \frac{\partial W_h}{\partial \Phi_h} + \frac{\partial \Phi_h}{\partial \Phi_h} + \frac{\partial \psi_h}{\partial \Phi_h}, \quad \frac{\partial n}{\partial \psi_h} = \frac{\partial W_h}{\partial \psi_h} + \frac{\partial \Phi_h}{\partial \psi_h} + \frac{\partial \psi_h}{\partial \psi_h}.$$

Observons que

$$\frac{\partial W_h}{\partial W_{h+1}} = -\frac{\partial W_h}{\partial W_{h+1}}, \quad \frac{\partial \Phi_h}{\partial \Phi_{h+1}} = -\frac{\partial \Phi_h}{\partial \Phi_{h+1}}, \quad \frac{\partial \psi_h}{\partial \psi_{h+1}} = -\frac{\partial \psi_h}{\partial \psi_{h+1}}$$

$$\frac{\partial W_h}{\partial \Phi_{h+1}} = -\frac{\partial W_h}{\partial \Phi_{h+1}}, \quad \frac{\partial \Phi_h}{\partial \psi_{h+1}} = -\frac{\partial \Phi_h}{\partial \psi_{h+1}}, \quad \frac{\partial \psi_h}{\partial \psi_{h+1}} = -\frac{\partial \psi_h}{\partial \psi_{h+1}}$$

$$\frac{\partial W_h}{\partial \psi_{h+1}} = -\frac{\partial W_h}{\partial \psi_{h+1}}, \quad \frac{\partial \Phi_h}{\partial \psi_{h+1}} = -\frac{\partial \Phi_h}{\partial \psi_{h+1}}, \quad \frac{\partial \psi_h}{\partial \psi_{h+1}} = -\frac{\partial \psi_h}{\partial \psi_{h+1}}$$

Ceci nous mène au résultat suivant, pour $h < L-1$,

$$\left\{ \left(\frac{\partial W_h}{\partial W_{h+1}} - \frac{\partial W_h}{\partial W_{h+1}} \right) + \left(\frac{\partial \Phi_h}{\partial \Phi_{h+1}} - \frac{\partial \Phi_h}{\partial \Phi_{h+1}} \right) + \left(\frac{\partial \psi_h}{\partial \psi_{h+1}} - \frac{\partial \psi_h}{\partial \psi_{h+1}} \right) \right\} \frac{\partial b_h}{\partial n} = f(b_h)$$

Parallèlement,

$$\left\{ N + \frac{\partial W_{L-1}}{\partial n} + \frac{\partial \Phi_{L-1}}{\partial n} + \frac{\partial \psi_{L-1}}{\partial n} \right\} \frac{\partial b_{L-1}}{\partial n} = f(b_{L-1})$$

Nous utilisons l'algorithme de Sethi (1963) pour résoudre $\partial n / \partial b_h = 0$. Il se fonde sur l'hypothèse que les dérivées partielles sont proportionnelles à des fonctions quadratiques en b_h . La valeur mise à jour de b_h est donnée par la racine

$n = 19$ à $n = 28$ vaut la peine d'être mentionnée. Pour les deux méthodes de répartition, le plan de sondage obtenu à l'aide du modèle log-linéaire produit des strates à tirage complet plus petites que celles obtenues par Lavallée et Hidiroglou.

Tableau 2

Plans d'échantillonnage stratifiés obtenus à l'aide de l'algorithme généralisé de Lavallée et Hidiroglou pour la population MU284 en utilisant REV84 comme variable de stratification, un modèle log-linéaire avec $\beta_{\log} = 1,1$ et $\sigma_{\log}^2 = 0,2116$ de la relation entre REV84 et RMT85, et un c.v. cible de 5 %

Algorithme de stratification à modèle log-linéaire avec répartition puissance où $p = 0,7$									
strate 1	strate 2	strate 3	strate 4	strate 5	b_h	mo	var	N_h	n_h f_h n
1 558	3 031	5 706	11 107	59 878					
1 023	2 219	4 022	7 602	25 536					
97 245	168 204	464 471	2 659 061	39 131 413	répartition optimum de Neyman				

à-dire distribution lognormale (voir Johnson et Kotz 1970), c'est-

$$E(e^y) = e^{\sigma^2_{\log}/2} \text{ et } \text{Var}(e^y) = e^{\sigma^2_{\log}}(e^{\sigma^2_{\log}} - 1).$$

Nous avons

$$E(Y|b_h \geq X > b_{h-1}) = \exp(\sigma^2_{\log}/2) E(X|b_{\log} | b_h \geq X > b_{h-1})$$

tandis que $\text{Var}(Y|b_h \geq X > b_{h-1})$ est égale à

$$\text{Var}(E(Y|X) | b_h \geq X > b_{h-1}) + E(\text{Var}(Y|X) | b_h \geq X > b_{h-1})$$

$$= \exp(2\alpha + \sigma^2_{\log}) \{ \text{Var}(X|b_{\log} | b_h \geq X > b_{h-1})$$

$$+ (e^{\sigma^2_{\log}} - 1) E(X|b_{\log} | b_h \geq X > b_{h-1}) \}$$

$$= \exp(2\alpha + \sigma^2_{\log}) \{ e^{\sigma^2_{\log}} E(X|b_{\log} | b_h \geq X > b_{h-1})$$

$$- E(X|b_{\log} | b_h \geq X > b_{h-1}) \}.$$

Dans certains cas, les valeurs des paramètres β_{\log} et σ^2_{\log} peuvent être calculées d'après des données historiques. Des

valeurs simples spéciales de ces paramètres sont $\beta_{\log} = 1$

et $\sigma^2_{\log} = (1 - p^2) \text{Var}(\log(X))$. Ici, p représente la corrélation supposée entre $\log(X)$ et $\log(X)$, à laquelle on peut

donner des valeurs prédéterminées, comme 0,95 ou 0,99.

3.2 Un modèle linéaire

Dans les textes traitant de l'échantillonnage, la relation entre Y et X est souvent modélisée au moyen d'un modèle linéaire hétéroscédastique,

$$Y = \beta_{\ln} X + e, \quad (3.5)$$

où la distribution conditionnelle de e , étant donné X , a une moyenne nulle et une variance $\sigma^2_{\ln} X$, pour un paramètre

donné non négatif γ . Des calculs simples donnent $E(Y|b_h \geq X > b_{h-1}) = \beta_{\ln} E(X|b_h \geq X > b_{h-1})$, tandis que

$$\text{Var}(Y|b_h \geq X > b_{h-1}) = \beta^2_{\ln} \{ \text{Var}(X|b_h \geq X > b_{h-1}) + (\sigma^2_{\ln}/\beta_{\ln})^2 E(X|b_h \geq X > b_{h-1}) \}.$$

tionnelle de Y dépend de trois moments conditionnels de X . La généralisation de l'algorithme de Sethi présentée à la section 5 ne marche pas dans cette situation. Notons, cependant que, si $\gamma = 2$, les moyenne et variance conditionnelles de Y sont proportionnelles à celles du modèle log-

linéaire avec

$$\beta_{\log} = 1 \text{ et } \sigma^2_{\log} = \log(1 + (\sigma^2_{\ln}/\beta_{\ln})^2); \quad (3.6)$$

les facteurs de proportionnalité sont $\exp(\alpha + \sigma^2_{\log}/2)/\beta_{\ln}$ et $\exp(2\alpha + \sigma^2_{\log})/\beta^2_{\ln}$ pour l'espérance conditionnelle et la variance conditionnelle, respectivement. Donc, les deux modèles de la relation entre la variable de stratification et la

variable étudiée, à savoir le modèle log-linéaire de la section 3.1 ou le modèle linéaire (3.5) avec le paramètre $\gamma = 2$, mènent, à la section 5, au même plan d'échantillonnage stratifié, à condition que les égalités (3.6) soient vérifiées. Plus loin, nous utiliserons le modèle log-linéaire pour représenter la relation entre X et Y . Ce modèle devrait donner de bons résultats quand la relation réelle entre Y et X est modélisée par (3.5) avec $\gamma \approx 2$. Si l'on suppose que le modèle (3.5) est vérifié pour une valeur plus faible de γ , l'algorithme de la section 5 reste applicable lorsqu'on fixe la valeur de γ à 0 ou à 1. Toutefois, nous ne nous attachons pas sur ce problème ici.

3.3 Un modèle à remplacement aléatoire

Ce modèle se fonde sur l'hypothèse que la variable de stratification est égale à la variable étudiée, c'est-à-dire $X = Y$, pour la plupart des unités. Il existe cependant une faible probabilité ε qu'une unité ait changé considérablement; le cas échéant, la valeur de Y est caractérisée par la fonction de densité $f(x)$ et est distribuée indépendamment de la valeur de X . Cette approche est celle utilisée dans Rivest (1999) pour modéliser l'occurrence des unités qui sautent d'une strate à une autre, unités pour lesquelles X n'est pas représentative de Y . Plus formellement, on peut écrire

$$Y = \begin{cases} X & \text{avec probabilité } 1 - \varepsilon \\ X^{\text{nov}} & \text{avec probabilité } \varepsilon \end{cases}$$

où X^{nov} représente une variable aléatoire dont la densité $f(x)$ est distribuée indépendamment de X . Sous ce modèle, la moyenne conditionnelle de Y est donnée par

$$E(Y|b_h \geq X > b_{h-1}) = (1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X),$$

tandis que sa variance conditionnelle est égale à

$$\text{Var}(Y|b_h \geq X > b_{h-1})$$

$$= (1 - \varepsilon) E(X^2|b_h \geq X > b_{h-1}) + \varepsilon E(X^2)$$

$$- \{ (1 - \varepsilon) E(X|b_h \geq X > b_{h-1}) + \varepsilon E(X) \}^2.$$

4. UN EXEMPLE

Avant de passer aux détails techniques de la construction des algorithmes, il serait utile de donner un exemple. Considérons la population MU284 de Särndal, Swensson et Wretman (1992) contenant des données sur 284 municipalités suédoises.

Pour élaborer un plan d'échantillonnage stratifié sur l'estimation de la moyenne de RMT85, c'est-à-dire les recettes provenant de l'imposition municipale de 1985, nous utilisons REV84, c'est-à-dire la valeur des biens

nous présentons l'application de l'algorithme de Sethi quand la variable de stratification et la variable étudiée diffèrent. Enfin, nous donnons des exemples numériques.

2. UNE REVUE DE L'ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

Dans la suite de l'article, nous utiliserons la notation générale relative à l'échantillonnage aléatoire stratifié suivante :

L = nombre de strates;

$W_h = N_h/N$ représente, pour $h = 1, \dots, L$, le poids relatif de la strate h , N_h représente la taille de la strate h et $N = \sum N_h$, la taille totale de la population;

n_h représente, pour $h = 1, \dots, L$, la taille d'échantillon dans la strate h et $f_h = n_h/N_h$ représente la fraction d'échantillonnage;

\bar{Y}_h et y_h représentent les moyennes de population et d'échantillon de Y dans la strate h ;

S_{y_h} représente l'écart-type de population de Y dans la strate h .

Dans le présent article, les strates sont créées en prenant X pour variable de stratification. La strate h comprend toutes les unités pour lesquelles la valeur de X est comprise dans l'intervalle $(b_{h-1}^{x^*}, b_h^{x^*}]$, où $-\infty = b_0 < b_1 < \dots < b_L = +\infty$ sont les limites de la strate.

Nous pouvons exprimer l'estimateur d'échantillon de Y sous la forme $\bar{y}_{st} = \sum W_h \bar{y}_h$; sa variance est donnée par :

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{1}{n_h} - \frac{1}{N} \right) S_{y_h}^2 \quad (2.1)$$

Dans les enquêtes auprès des entreprises, toutes les grandes entreprises sont sélectionnées; nous posons que la strate L est la strate à tirage complet, de sorte que $n_L = N_L$. Si $h < L$, n_h on peut écrire la taille de l'échantillon dans la strate à tirage partiel h sous la forme $(n - N_L) a_h$, où n est la taille totale de l'échantillon et a_h dépend de la règle de répartition. Les deux règles de répartition envisagées ici sont :

— la règle de répartition par la méthode puissance

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^L (W_k \bar{Y}_k)^p} \quad (2.2)$$

où p est un nombre positif compris dans $(0, 1]$;

— la règle de répartition optimum de Neyman

Les limites optimales de strate sont les valeurs de b_1^*, \dots, b_{L-1}^* qui minimisent n , sous une condition de précision de \bar{y}_{st} telle que $\text{Var}(\bar{y}_{st}) = Y^2 c^2$, où c est le coefficient de variation (c.v.) cible. Pour les enquêtes auprès des entreprises, on utilise souvent la fourchette $c = 1\% \text{ à } 10\%$.

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{y_h}^2 / a_h}{\text{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h^2 S_{y_h}^2 / N} \quad (2.4)$$

La résolution de (2.1) pour trouver la valeur de n donne

$$a_h = \frac{W_h S_{y_h}}{\sum_{k=1}^{L-1} W_k S_{y_k}} \quad (2.3)$$

Dans la suite, il est pratique de considérer X et Y comme des variables aléatoires continues et de représenter par $f(x)$, $x \in R$ la densité de X . Les données $\{x_i, i = 1, \dots, N\}$ peuvent être considérées comme N réalisations indépendantes de la variable aléatoire X . Puisque la strate h comprend les unités de population pour lesquelles la valeur de X est comprise dans l'intervalle $(b_{h-1}^{x^*}, b_h^{x^*}]$, la stratification est basée sur les valeurs de $E[X|b_h^{x^*} \geq X > b_{h-1}^{x^*}]$ et $\text{Var}[X|b_h^{x^*} \geq X > b_{h-1}^{x^*}]$, c'est-à-dire la moyenne et la variance conditionnelles de X , étant donné que l'unité est comprise dans la strate h , pour $h = 1, \dots, L-1$. Suivent trois modèles de la relation entre X et Y , ainsi que leurs moyennes et variance conditionnelles respectives pour X .

3.1 Un modèle log-linéaire

Notre premier modèle a la forme $\log(Y) = \alpha + \beta_{\log}(X) + \epsilon$, où ϵ est une variable aléatoire normale de moyenne nulle et de variance σ_{\log}^2 qui est indépendante de X , et α et β_{\log} sont des paramètres qu'il faut déterminer. Si $\alpha = 0$, $\beta_{\log} = 1$ et $\sigma_{\log}^2 = 0$, on a $X = Y$; la variable étudiée et la variable de stratification sont identiques. En général, $Y = e^{\alpha} X^{\beta_{\log}}$. On peut évaluer les moments conditionnels de Y d'après les propriétés fondamentales de la loi de

Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises

LOUIS-PAUL RIVEST¹

RÉSUMÉ

Le présent article décrit des algorithmes de stratification permettant de tenir compte d'une divergence entre la variable de stratification et la variable étudiée lors de l'élaboration d'un plan de sondage stratifié. Nous proposons deux modèles pour caractériser la relation entre ces deux variables. L'un est un modèle de régression log-linéaire; l'autre suppose que la variable étudiée et la variable de stratification coïncident pour la plupart des unités, mais que des divergences importantes existent pour certaines unités. Puis, nous modifions l'algorithme de stratification de Lavallée et Hidiroglou (1988) afin d'intégrer ces modèles dans la détermination des tailles d'échantillon et des limites de strates optimales pour un plan de sondage stratifié. Ensuite, nous donnons un exemple pour illustrer la performance du nouvel algorithme de stratification. Enfin, nous décrivons l'application numérique de cet algorithme.

MOTS CLÉS : Répartition optimum de Neyman; répartition par la méthode puissance; échantillonnage aléatoire stratifié.

1. INTRODUCTION

L'élaboration de plans d'échantillonnage stratifiés par les statisticiens ne date pas d'aujourd'hui. Dans Cochran (1977), les chapitres 5 et 5A sont consacrés à l'examen de plusieurs méthodes de répartition d'une population en strates. La création de strates est une question abordée couramment dans les publications statistiques. Les contributions récentes incluent Hedlin (2000), qui réexamine la règle de stratification d'Ekman (1959), et Dorfman et Valiant (2000), qui comparent la stratification basée sur un modèle à l'échantillonnage équilibré. La stratification basée sur un modèle fait l'objet d'une discussion dans Godfrey, Särndal, Swensson et Wretman (1992). Les populations visées par les enquêtes auprès des entreprises ont une distribution asymétrique; un petit nombre d'unités représentent une part importante du total de la variable étudiée. Par conséquent, il convient d'inclure toutes les grandes unités dans l'échantillon (Dalenius 1952; Glasser 1962). Un bon plan d'échantillonnage comprend une strate à tirage complet pour les grandes entreprises, où toutes les unités sont échantillonnées, ainsi que des strates à tirage partiel pour les petites et moyennes entreprises. Habituellement, la fraction d'échantillonnage diminue parallèlement à la taille de l'unité; des poids d'échantillonage importants sont appliqués aux petites entreprises. L'algorithme de stratification de Lavallée et Hidiroglou (1988) est souvent utilisé pour déterminer les limites de strate et les tailles d'échantillon de strate dans ce contexte (consulter, par exemple, Plana et Krenzke 1994, 1996). Cet algorithme comprend une variable de stratification, que l'on connaît pour toutes les unités de la population. Il donne les limites de strate et les tailles d'échantillon de strate qui

réduisent au minimum la taille totale de l'échantillon nécessaire pour atteindre le niveau voulu de précision. Il est basé sur une méthode itérative, élaborée par Sethi (1963) pour déterminer les limites optimales de strate. L'algorithme de Lavallée et Hidiroglou ne tient pas compte des différences éventuelles entre la variable de stratification et la variable étudiée. À mesure que le temps passe, cette différence augmente et le plan d'échantillonnage produit par cet algorithme risque de ne plus satisfaire les critères de précision. Dalenius et Gurney (1951), ainsi que Cochran (1977, chapitre 5A) considèrent la stratification dans les situations où la variable étudiée et la variable de stratification diffèrent. Nombre d'auteurs ont étudié des formules appropriatives pour déterminer les limites de strate et pour évaluer le gain de précision dû à la stratification sur une variable auxiliaire. À cet égard, Serfling (1968), Singh et Sukatne (1969), Singh (1971), Singh et Parkash (1975), Anderson, Kish et Corneli (1976), Oslo (1976), Wang et Aggarwal (1984) et Yavada et Singh (1984) sont des contributions pertinentes. Hidiroglou et Srinath (1993) et Hidiroglou (1994) proposent des méthodes de mise à jour des limites de strate basées sur une nouvelle variable de stratification. Cependant, ces articles ne fournissent explicitement aucun algorithme de stratification tenant compte de la divergence entre la variable de stratification et la variable étudiée. Le présent article comble cette lacune grâce à la construction de généralisations de l'algorithme de Lavallée et Hidiroglou (1988) qui expriment la relation entre ces deux variables sous forme de modèle statistique. Nous passons d'abord brièvement en revue l'échantillonnage stratifié et les méthodes de répartition d'échantillon. Puis, nous proposons des modèles de la relation à la variable de stratification et la variable étudiée. Ensuite,

¹ Louis-Paul Rivest, Département de mathématiques et de statistique, Université Laval, Ste-Foy, (Québec) Canada, G1K 7P4.

où

$$\sigma^2(P) = P(1 - P), \text{ et } E_2^{NS1*} = \left(\frac{N}{X}\right)^2.$$

BIBLIOGRAPHIE

COHEN, S.B. (1998). Sample design of the 1996 medical expenditure panel survey medical provider component. *Journal of Economic and Social Measurement*, 24, 25-53.

DICKER, M., et SUNSHINE, J.H. (1987). Family use of health care, United States, 1980. *National Health Care Utilization and Expenditure Survey*. Rapport No. 10. DHHS Pub. 87-20210.

K. et WAKSBERG, J. (1995) National Health Care Survey: List versus Network Sampling. Rapport non-publié. National Center for Health Statistics.

JUDKINS, D., MARKER, D., WAKSBERG, J., BOTMAN, S. et MASSEY, J. (1999). National Health Interview Survey: Research for the 1995-2004 redesign. National Center for Health Statistics. *Vital and Health Statistics*. Washington, DC: Government Printing Office Series 2, 126, 76-89.

KISH, L. (1982). Design effect. *Encyclopedia of the Statistical Sciences*. John Wiley & Sons, Inc. 2, 347-348.

LEAVER, S., et VALLIANT, R. (1995). Statistical problems in estimating the U.S. consumer price index. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christanson, M.J. Colledge et P.S. Kott). New York: John Wiley & Sons, Inc.

MASSEY, L.T., MOORE, T.F., PARSONS, V. et TADRO, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics. *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2, 110.

SIRKEN, M., et SHIMIZU, I. (1999). Enquêtes auprès des établissements fondées sur un échantillon représentatif de ménages: L'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 25, 213-218.

SIRKEN, M., SHIMIZU, I. et JUDKINS, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1, 470-473.

SIRKEN, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics*. John Wiley & Sons, Inc. 4, 2977-2986.

SIRKEN, M.G. (2001). The Hansen-Hurwitz estimator revisited: sampling without replacement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Sous presse.

THOMPSON, S. (1992). *Sampling*. New York: John Wiley & Sons, Inc. 117-118.

WUNDERLICH, G.S. (Ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. Washington, DC: National Academy Press.

$$\sigma_2^{NS1*}(P) = P\sigma_2^{NS1*} + \sigma^2(P)E_2^{NS1*}, \quad 0 < P \leq 1$$

se décompose en deux parties

$$\sigma_2^{NS1*} = \frac{1}{N} \sum_{i=1}^N \sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{N}{X} \left(\frac{N}{X}\right)^2$$

$$P = \frac{N}{N^*}$$

$$\sigma_2^{NS1*} = \frac{1}{N^*} \sum_{i=1}^N \sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{N}{X} \left(\frac{N}{X}\right)^2$$

est la variance de la population tronquée de premier degré de l'estimateur NS, à l'exclusion des $N_0 = N - N^*$ ménages ne comptant aucune transaction avec les établissements,

$$\sigma^2(P) = P(1 - P)$$

est la variance de la variable binomiale P , et

$$E_2^{NS1*} = (X/N)^2$$

est le carré de l'espérance de la variable x répartie sur les N^* ménages.

Preuve

Ajoutons X/N^* au premier terme du deuxième membre de (A.1) et soustrayons l'en.

$$\frac{1}{N^*} \sum_{i=1}^N \sum_{j=1}^R M_{ij} \bar{X}_j - \frac{n}{X} \left(\frac{N}{X}\right)^2 = \frac{P}{N^*} \sum_{i=1}^N \sum_{j \in A_i} \left(M_{ij} \bar{X}_j - \frac{N}{X} \right)^2 + P \left(\frac{N}{X} \right)^2$$

$$= P\sigma_2^{NS1*}(P) + P \left(\frac{N}{X} \right)^2 \quad (A.2)$$

Remplaçons le premier terme du deuxième membre de (A.1) par (A.2).

$$\sigma_2^{NS1*}(P) = P\sigma_2^{NS1*} + P \left(\frac{N}{X} \right)^2 + (1 - P) \left(\frac{N}{X} \right)^2 \quad (A.3)$$

$$= P\sigma_2^{NS1*}(P) + \sigma^2(P)E_2^{NS1*}$$

Les estimateurs NS et HH sont aussi efficaces l'un que l'autre si, et uniquement si, chaque ménage de la population est inclus dans une seule transaction. Sinon, ni l'estimateur NS ni l'estimateur HH n'est nécessairement plus efficace que l'autre. Néanmoins, il semble probable que l'estimateur HH soit plus efficace que l'estimateur NS en cas d'échantillonnage à un seul degré des établissements, voire même considérablement plus efficace, particulièrement lorsque des fractions importantes de ménages ne font aucune transaction et (ou) lorsque le regroupement des transactions à l'intérieur des ménages est dû principalement aux ménages qui font plusieurs transactions avec le même établissement plutôt qu'aux ménages qui font des transactions avec plusieurs établissements. Dans le cas de l'échantillonnage à deux degrés, le résultat n'est pas aussi évident que dans le cas de l'échantillonnage à un degré, car la composante de l'estimateur HH excède de la variance de l'estimateur NS d'une valeur qui dépend de celle de l'estimateur NS d'une valeur qui dépend de l'importance du regroupement des transactions multiples avec un même établissement à l'intérieur des ménages. On pourrait soutenir que la principale limite du modèle d'erreur présentée ici est la présomption que les bases de sondage autonome et établie d'après une enquête de démographie ne comportent aucune erreur de couverture ni de mesure de taille. Toutefois, les coûts comparatifs de la création et de la tenue d'une liste d'établissements autonome ou établie d'après une enquête démographique varient considérablement. Une enquête à l'autre. Quoique le modèle vise à équilibrer les coûts des enquêtes auprès des établissements fondées sur chaque catégorie de bases de sondage, il ne tient pas compte des coûts différents de création et de tenue à jour de chacune des catégories de bases de sondage.

Même si l'on ne dispose pas de données empiriques sur les coûts comparatifs de la création et de la tenue à jour des bases de sondage, il est juste de dire que la base de sondage établie d'après une enquête démographique devrait être considérée sérieusement comme plan de sondage de base de sondage indépendant lorsque la création et la tenue à jour de bases de sondage indépendantes de bonne qualité est impossible, exorbitante ou demande trop de temps, et (ou) lorsque la création et la tenue à jour de listes d'établissements établies d'après une enquête démographique est relativement peu coûteuse. Par exemple, la base de sondage établie d'après une enquête démographique serait particulièrement intéressante comme remplacement éventuel de la base de sondage indépendante lorsque cette dernière est difficile à créer et à tenir à jour à cause de changements rapides dus à la création, à la fermeture et à la fusion d'établissements, et que le coût de la base de sondage établie d'après une enquête démographique est assez faible parce qu'elle peut être créée et tenue à jour sous forme de sous-produit d'une enquête démographique par sondage en cours (Wunderlich 1992) et (ou) sous forme de sous-produit d'un programme permanent d'appariement des transactions des ménages recensés lors d'une enquête

Si on l'exprime sous forme de fonction de P , c'est-à-dire la traction de ménages comptant au moins une transaction, la variance de population de premier degré de l'estimateur par échantillonnage en réseau (NS) de X

ANNEXE

L'auteur remercie les deux évaluateurs et, surtout, un rédacteur adjoint, pour leurs commentaires très constructifs. Les opinions exprimées dans le présent article n'engagent que l'auteur et ne représentent pas nécessairement les vues ou positions officielles du National Center for Health Statistics.

REMERCIEMENTS

L'auteur remercie les deux évaluateurs et, surtout, un rédacteur adjoint, pour leurs commentaires très constructifs. Les opinions exprimées dans le présent article n'engagent que l'auteur et ne représentent pas nécessairement les vues ou positions officielles du National Center for Health Statistics.

L'auteur remercie les deux évaluateurs et, surtout, un rédacteur adjoint, pour leurs commentaires très constructifs. Les opinions exprimées dans le présent article n'engagent que l'auteur et ne représentent pas nécessairement les vues ou positions officielles du National Center for Health Statistics.

Une autre limite du modèle tient à l'hypothèse irréaliste selon laquelle l'enquête démographique a partir de laquelle est produite la liste d'établissements est fondée sur un plan d'échantillonnage à un degré en vertu duquel les ménages sont sélectionnés avec probabilité égale et avec remise. En fait, les enquêtes démographiques sont presque toujours fondées sur un plan d'échantillonnage à plusieurs degrés selon lequel les ménages sont sélectionnés sans remise à l'étape finale d'échantillonnage. Habituellement, l'hypothèse que la sélection se fait par EAS à tendance à produire une estimation considérablement sous-estimée de la variance de l'estimateur NS et, par conséquent, à pour effet d'exagérer l'efficacité relative de cet estimateur comparativement à l'estimateur HH. Par contre, l'hypothèse selon laquelle les ménages sont échantillonnés avec remise à effets opposés, mais ceux-ci sont modérés (Sjorken 2001) comparativement à ceux de l'hypothèse d'EAS. Toutefois, le modèle d'erreur peut être appliqué aux autres plans d'échantillonnage, non considérés ici, que l'on utilise pour réaliser les enquêtes démographiques.

$\sigma_j^2 = 0$ ($j=1, 2, \dots, R$). Cependant, à part ces contingences, la variance de deuxième degré est toujours plus importante pour l'estimateur HH que pour l'estimateur NS, et la

grandeur de la différence dépend de l'importance du regroupement des transactions avec un même établissement

à l'intérieur des ménages, et de l'importance des variances à l'intérieur des établissements

Si aucun des N^* ménages ne fait plusieurs transactions

avec le même élablissement, la différence entre les variances des estimateurs HH et NS est la même pour les

enquêtes sur échantillon d'établissements à un seul degré et à deux degrés. Sinon, la différence entre les variances HH

et NS est plus faible pour l'échantillonnage à deux degrés

que celui d'un seul degré des établissements, car, quand les ménages font des transactions multiples avec le même

établissemment, la variance de second degré est plus importante pour l'estimateur HH que pour l'estimateur NS.

5. RÉSUMÉ ET CONCLUSION

d'échantillonnage à deux degrés

d'enquêtes sur échantillons à un degré et à deux degrés auprès des établissements. L'estimateur de Hansen-Hurwitz

(HH) est fondé sur une base de sondage autonome énu-

transactions avec tous les ménages durant une période

particulière de l'année civile. L'estimateur par échantillon-
nage en réseau (NS pour *network sampling*) dépend d'une

$$(22) \quad \frac{{}^f\mathcal{W}}{{}^f\mathcal{D}} \sum_R^{\mathcal{I}=\mathcal{I}} \frac{u}{N} = \left\{ \frac{{}^f\mathcal{W}}{{}^{(\mathcal{I})}\mathcal{W}^{1-\mathcal{I}}\mathcal{W}^{\mathcal{I}}\mathcal{W}} \sum_N^{\mathcal{I}=\mathcal{I}} - (1-{}^f\mathcal{W}) \right\} {}^f\omega \sum_R^{\mathcal{I}=\mathcal{I}} \frac{\mathcal{I}u}{N}$$

plans de sondage différents. En cas d'échantillonnage à un seul degré, l'estimateur HH dépend d'un plan de sondage

où les établissements sont les unités d'échantillonnage sélectionnées avec PBT et avec remise et l'estimateur NS

dépend d'un plan de sondage où les ménages sont les unités

d'échantillonnage que l'on sélectionne avec la même probabilité que dans l'enquête démographique, enquête que l'on

suppose, dans le modèle d'erreur, être réalisée par EAS

degrés, les transactions sont les unités d'échantillonnage de

deuxième degré des estimateurs HH et NS. L'estimateur HH est basé sur des échantillons de transactions de taille

fixe qui sont sélectionnées par EAS indépendamment, sans remise. L'estimateur NS est basé sur des échantillons de

transactions dont la taille est proportionnelle au nombre de

ment et qui sont sélectionnées indépendamment par EAS

...sans remise.

Soit $P = \frac{N^*}{N}$ = fraction de N ménages comptant au moins une transaction,

$P_0 = 1 - P = \frac{N}{N^*}$ = fraction de N ménages sans aucune transaction.

Nous démontrons à l'annexe que la variance de population à un seul degré de l'estimateur NS de X , si on l'exprime sous forme de fonction de P , se décompose en deux parties

$$\sigma_{NS1}^2(P) = \frac{N}{N^*} \sum_{i=1}^I \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{N}{X} \right)^2 + \sigma^2(P) E_2^{NS1}, \quad 0 < P \leq 1 \quad (12)$$

où

$$\sigma_{NS1}^2 = \frac{1}{N^*} \sum_{i=1}^I \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{N}{X} \right)^2 \quad (13)$$

représente la variance de population à un seul degré de la variable x sur la population tronquée de N^* ménages

$$E_2^{NS1} = \left(\frac{N}{X} \right)^2 = \frac{1}{N^*} \sum_{i=1}^I \left(\sum_{j \in A_i} M_{ij} \bar{X}_j \right)^2 - \sigma_{NS1}^2 \quad (14)$$

représente le carré de l'espérance de la variable x sur la population tronquée de N^* ménages et

$$\sigma^2(P) = P(1 - P) \quad (15)$$

est la variance de la variable binomiale P . Pour une valeur fixe de M , la fonction $\sigma_{NS1}^2(P|M)$ est maximale quand

$$P = P_{\max} = \frac{1}{2} \left[\left(\sigma_{NS1}^2 / E_2^{NS1} \right) + 1 \right] \leq 1.$$

Si $\sigma_{NS1}^2 \geq E_2^{NS1}$, $P_{\max} = 1$ et si $\sigma_{NS1}^2 < E_2^{NS1}$, $1/2 < P_{\max} < 1$. Quand $P = 1$, $\sigma^2(P) = 0$ et, par conséquent, $\sigma_{NS1}^2(P) = \sigma_{NS1}^2$. Si $P = \bar{M} = (M/N) = 1$, ce qui sous-entend que chacun des N ménages compte une seule transaction,

$$\sigma_{NS1}^2(P = \bar{M} = 1) = \sigma_{NS1}^2(N^* = M) = \sigma_{HH1}^2 \quad (16)$$

car

$$\sigma_{NS1}^2(N^* = M) = \frac{1}{N^*} \sum_{i=1}^I \sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{N}{X} \left(\sum_{j=1}^J \frac{M}{X} \bar{X}_j - \frac{M}{X} \right)^2 = \sigma_{HH1}^2 \quad (17)$$

remise.

et $\sigma^2(P = 1) = 0$. Autrement dit, si $P = 1$, ce qui sous-entend que chacun des N ménages compte une seule transaction, la variance de l'estimateur NS₁, qui dépendrait alors de l'EAS de transactions avec remise, est équivalente à la variance de l'estimateur HH₁, qui dépend d'un échantillon en grappe PPT de taille équivalente sélectionné avec

4.3 Effet de plan de sondage en cas d'échantillonnage à un seul degré

Soit $X_{NS1} = \frac{n}{N} \sum_{i=1}^I \sum_{j \in A_i} M_{ij} \bar{X}_j$ = l'estimateur NS non biaisé de X en cas d'échantillonnage à un seul degré,

$$X_{HH1} = \frac{M}{\sum_{j=1}^J \bar{X}_j} = \frac{R}{\sum_{j=1}^J \bar{X}_j} = \text{l'estimateur HH non biaisé de } X \text{ en cas d'échantillonnage à un seul degré.}$$

en cas d'échantillonnage à un seul degré.

Définissons l'effet de plan de sondage total de l'échantillonnage à un seul degré pour l'estimateur NS₁ comme étant le rapport des variances des estimateurs NS₁ et HH₁ pour des tailles équivalentes d'échantillons conditionnelles sur l'ensemble des échantillons de n ménages.

$$\lambda(P) = \frac{\text{Var}(X_{NS1})}{\text{Var}(X_{HH1})} = \frac{\bar{M}^2}{\sigma_{HH1}^2} \quad (18)$$

où $\lambda(P) < 1$ indique que l'estimateur NS₁ est plus efficace que l'estimateur HH₁, et $\lambda(P) > 1$ indique que l'estimateur HH₁ est plus efficace que l'estimateur NS₁. Nous avons noté dans (12) et (15) que $\sigma_{NS1}^2(P) = P \sigma_{NS1}^2 + P(1 - P)(X/N)^2$, et dans (16), que $\sigma_{HH1}^2 = \sigma_{NS1}^2(N^* = M)$. Si nous faisons ces substitutions dans (18), l'effet total de plan de sondage devient

$$\lambda(P) = \text{def}_{NS1}^2 + (1 - P) Z_{NS1}, \quad 0 < P \leq 1 \quad (19)$$

où

$$Z_{NS1} = \frac{\bar{M}^2 \sigma_{NS1}^2}{P(X/N)^2} \quad (20)$$

est l'effet attribuable aux N_0 ménages ne comptant aucune transaction, et

$$\text{def}_{NS1}^2 = \frac{\left[\frac{\bar{M}^2}{\sigma_{NS1}^2} \right]}{\left[\frac{\bar{M}^2}{\sigma_{NS1}^2} \right]} = \frac{\bar{M}^2 \sigma_{NS1}^2}{P \sigma_{NS1}^2} \quad (21)$$

est l'effet attribuable aux N^* ménages comptant des transactions. Autrement dit, def_{NS1}^2 est l'effet de plan de sondage de l'échantillonnage en réseau d'une population de N^* grappes de ménages comptant au moins une transaction, avec probabilités égales et remise, comparativement à l'échantillonnage en réseau d'une population de M

sans remise pour chaque établissement f lié au ménage H_i , la variance de l'estimateur NS (5) est donnée par (Sirken et coll., 1995) :

$$\text{Var}(X'_{iNS}) = \frac{n}{N^2} \sigma_{NS1}^2 + \frac{n}{N} \sum_{j=1}^{N_{NS}} \sum_{R} M_{ij} \frac{M_{ij} - t_{NS} M_{ij}}{M_{ij}^2} \sigma_f^2 \quad (7)$$

où les premier et deuxième termes du deuxième membre représentent, respectivement, les composantes de premier et de deuxième degré de la variance,

$$\sigma_{NS1}^2 = \frac{1}{N} \sum_{j=1}^{N_{NS}} \left(\sum_{f \in A_j} M_{ij} \bar{X}_f - \bar{X} \right)^2 \quad (8)$$

est la variance de population entre les ménages, et σ_f^2 la variance de population à l'intérieur de l'établissement f telle que définie en (4). Une estimation non biaisée de la variance NS est donnée par

$$\text{Var}(X'_{iNS}) = \frac{n}{N^2} \left[\sum_{j=1}^{N_{NS}} \sum_{f \in A_j} M_{ij} \bar{X}_f (i) - \bar{X} \right]^2 \quad (9)$$

$$\text{où } \bar{X}' = X'/N.$$

4. LE MODÈLE D'ERREUR

4.1. Variances HH et NS pour des tailles prévues d'échantillon équivalentes

Obtenu en soustrayant (2) de (7), la différence entre les variances des estimateurs HH et NS de X est

$$\begin{aligned} \text{Var}(X'_{iNS}) - \text{Var}(X'_{iHH}) &= \left[\frac{n}{N^2} \sigma_{NS1}^2 - \frac{r}{M^2} \sigma_{HH1}^2 \right] + \left[\frac{n}{N} \sum_{j=1}^{N_{NS}} \sum_{R} M_{ij} \frac{M_{ij} - t_{NS} M_{ij}}{M_{ij}^2} \sigma_f^2 \right. \\ &\quad \left. - \frac{r}{M} \sum_{R} M_{ij}^{HH} (M_{ij} - t_{HH}) \sigma_f^2 \right] \quad (10) \end{aligned}$$

où les premier et deuxième ensembles de termes entre crochets dans le deuxième membre représentent, respectivement, les différences entre les composantes de premier et de deuxième degré de la variance des estimateurs HH et NS de X .

Représentons par $m_{NS} = r t_{NS}$ la taille de l'échantillon de transactions dans l'enquête auprès des établissements fondée sur la base de sondage établie d'après une enquête démographique, où t_{NS} , un entier positif, est la taille de l'échantillon de transactions sélectionné par transaction des ménages échantillonnés, et par $r = \sum_{i=1}^N \sum_{f \in A_i} M_{ij}$, la somme des transactions sur n ménages échantillonnés.

Manifestement, t est une variable aléatoire et son espérance conditionnelle sur l'ensemble des échantillons d'échantillon de transactions de l'estimateur NS conditionnelle sur l'ensemble des échantillons de n ménages. Représentons par $m_{HH} = r t_{HH}$ la taille de l'échantillon de transactions dans le cas de l'enquête auprès des établissements fondée sur la base de sondage autonome, où r est la taille de l'échantillon d'établissements et t_{HH} la taille de l'échantillon de transactions par établissement sélectionné. Posons que $r = E(t|n) = n\bar{M}$ et $t_{HH} = t_{NS} = t$, et il s'ensuit que les tailles prévues des échantillons de transactions des établissements NS et HH conditionnelles sur l'ensemble des échantillons de n ménages sont équivalentes, à savoir $E(m_{HH}|n) = tE(t|n) = n\bar{M} = E(m_{NS}|n)$.

Un tel calage des tailles des échantillons d'établissements et de transactions assure que les enquêtes HH et NS mènent à des listes d'établissements produites d'après une création et de tenue à jour des listes d'établissements autonome et des listes d'établissements produites d'après une enquête démographique.

Si l'on substitue $r = n\bar{M}$, $t_{HH} = t_{NS} = t$ et $M = N\bar{M}$ dans la formule (9), la différence entre les variances NS et HH pour des tailles prévues équivalentes d'échantillons d'établissements et de transactions conditionnelles sur l'ensemble des échantillons de n de ménages est

$$\text{Var}(X'_{iNS}) - \text{Var}(X'_{iHH}) = \frac{n}{N^2} [\sigma_{NS1}^2 - \bar{M} \sigma_{HH1}^2] - \frac{n}{N} \sum_{j=1}^{N_{NS}} \sum_{R} \frac{M_{ij} (M_{ij} - M_{ij})}{M_{ij}^2} \sigma_f^2 \quad (11)$$

Les premier et deuxième termes du deuxième membre de l'équation (11) représentent, respectivement, la différence entre les composantes de premier et de deuxième degrés des variances des estimateurs NS et HH pour des tailles prévues équivalentes d'échantillons conditionnelles sur l'ensemble des échantillons de n ménages.

4.2. Décomposition de la variance de population NS à un seul degré

Habituellement, certains ménages ne font de transaction avec aucun établissement et la proportion varie selon le type d'établissement, certains ménages ne font de transaction avec aucun établissement et la proportion varie selon le type de soins médicaux par les familles varie fortement selon la catégorie de fournisseurs de soins de santé (Dicker et Sunshine 1987). Pour une période de 12 mois, 70 % de familles n'ont pas déclaré d'hospitalisation, 7 % n'ont pas déclaré de visite à un médecin en consultation externe et 28 % n'ont pas déclaré de visite chez un dentiste.

L'établissement f ($f = 1, 2, \dots, R$) est égale à $t_{NS}^{\sum_{i=1}^n M_{ij}^{NS}}$ et la taille totale de l'échantillon de transactions est égale à $t_{NS}^{\sum_{i=1}^n \sum_{j \in A_i} M_{ij}^{NS}}$, c'est-à-dire la somme des transactions sur n ménages échantillonnés, est une variable aléatoire.

L'estimateur NS de X est

$$X_{NS}^{NS} = \sum_{i=1}^n \frac{1}{n} \sum_{j \in A_i} \pi_i M_{ij} X_j^f(i)$$

où A_i représente la grappe d'établissements distincts qui font des transactions avec le ménage échantillonné H_i , et

$$\bar{X}_{jk}^{NS} = \sum_{i=1}^n \sum_{j \in A_i} X_{jk}^f(i) / (t_{NS} M_{ij}^{NS})$$

est une estimation non biaisée \bar{X}_j pour un échantillon de $t_{NS} M_{ij}^{NS}$ transactions de l'établissement j . Comme les ménages sont sélectionnés avec remise, l'estimateur NS compte la quantité $\sum_{j \in A_i} M_{ij} X_j^f(i)$ chaque fois que le ménage H_i ($i = 1, 2, \dots, n$) est sélectionné dans l'échantillon et, comme un même établissement fait des transactions avec plusieurs ménages, l'estimateur NS compte la quantité $M_{ij} X_j^f(i)$ chaque fois qu'un ménage échantillonné fait des transactions avec l'établissement j ($j = 1, 2, \dots, n$) fait des transactions avec l'établissement j .

Si l'on suppose que l'enquête démographique est réalisée par BAS, $\pi_i = n/N$, et l'estimateur basé sur l'échantillonnage en réseau est

$$X_{NS}^{NS} = \sum_{i=1}^n \sum_{j \in A_i} M_{ij} X_j^f(i) / n$$

L'estimateur NS est un estimateur non biaisé de X .

$$E(X_{NS}^{NS}) = \sum_{i=1}^n \sum_{j \in A_i} E(M_{ij} X_j^f(i)) = \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}_j^f(i) = \sum_{i=1}^n \sum_{j \in A_i} M_{ij} X_j = X$$

L'estimateur NS représenté par l'équation (5) est autopondéré, parce que nous avons supposé que les n ménages sont sélectionnés par BAS. L'estimateur sera autopondéré si le plan d'échantillonnage de l'enquête démographique par sondage utilisée pour établir la liste d'établissements est autopondéré. Si $N^* = N$, ce qui sous-entend que N^* ménages ne font chacun qu'une seule transaction et que $N_0 = N - N^*$ ménages ne font aucune transaction, et si $n = t_{NS} = t_{HH}$, les estimateurs HH et NS sont équivalents.

$$X_{NS}^{NS} = \frac{n}{N} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} X_j^f(i) = \frac{n}{N} \sum_{i=1}^n \sum_{j \in A_i} \frac{1}{R} \sum_{f=1}^R X_j^f = \frac{r}{R} \sum_{f=1}^R \bar{X}_j^f = X_{HH}^{HH}$$

Dans les conditions d'EAS avec remise de n ménages et de sélection indépendante de $t_{NS} M_{ij}^{NS}$ transactions par BAS

L'estimateur HH non biaisé autopondéré PPT de X est

$$X_{HH}^{HH} = \frac{r}{r} \sum_{f=1}^R \bar{X}_j^f$$

où $\bar{X}_j^f = \sum_{i=1}^{t_{HH}} X_{ij}^f / t_{HH}$ est l'estimation non biaisée de $\bar{X}_j = X_j / M_j$ ($j = 1, 2, \dots, R$). Comme les établissements sont sélectionnés avec remise, l'estimateur HH compte \bar{X}_j^f autant de fois que l'établissement j est sélectionné dans l'échantillon.

La variance de X_{HH}^{HH} est (Thompson 1992)

$$\text{Var}(X_{HH}^{HH}) = \frac{r}{M^2} \sigma_{HH}^2 + \frac{r}{M} \sum_{f=1}^R (M_j - t_{HH}) \sigma_j^2 \quad (2)$$

où les premier et deuxième termes du deuxième membre représentent, respectivement, les composantes de premier et de deuxième degré de la variance,

$$\sigma_{HH}^2 = \frac{1}{R} \sum_{j=1}^R M_j (\bar{X}_j - X/M)^2 \quad (3)$$

est la variance entre établissements, et

$$\sigma_j^2 = \frac{M_j}{1} \sum_{k=1}^R (X_{jk}^f - X_j/M_j)^2 \quad (4)$$

est la variance à l'intérieur de l'établissement j .

3.2 Estimateur et variance NS

Considérons une enquête par échantillonnage à deux degrés auprès des établissements fondée sur une base de sondage établie d'après une enquête démographique. La base de sondage est une liste de n ménages échantillonnés H_i ($i = 1, 2, \dots, n$) qui ont participé à une enquête démographique par sondage. Pour chaque ménage énuméré H_i , la base fournit π_i , c'est-à-dire la probabilité de sélection dans l'enquête-ménage et M_{ij} , c'est-à-dire le nombre de transactions du ménage avec chaque établissement j ($j = 1, 2, \dots, R$). (Les M_{ij} sont déclarés par les membres répondants des ménages lors de l'enquête démographique par sondage).

Chacun des n ménages figurant sur la liste établie d'après l'enquête démographique représente une grappe d'établissements dont la taille varie de 0 à R établissements avec lesquels le ménage a fait des transactions. Les n grappes d'établissements représentent les unités primaires d'échantillonnage et les M_j ($j = 1, 2, \dots, R$) transactions des établissements échantillonnés représentent les unités d'échantillonnage secondaires. On sélectionne l'échantillon de transactions pour l'établissement j ($j = 1, 2, \dots, R$) de la façon suivante : un échantillon aléatoire simple de transactions de taille $t_{NS} M_j^f > \text{Min}(M_j, M_j^f)$ est sélectionné indépendamment sans remise pour chaque ménage échantillonné H_i ($i = 1, 2, \dots, n$), où t_{NS} est un entier positif. La taille de l'échantillon de transactions de

de l'estimateur NS à deux degrés est subdivisée en composante de la variance représentant les effets des ménages qui font et qui ne font pas de transactions, et la section 4.3 montre les effets de plan de sondage de l'estimateur NS en cas d'échantillonnage à un seul degré. Les composantes de deuxième degré de la variance des estimateurs NS et HH sont comparées à la section 4.4. Pour conclure, la section 5 résume les principaux résultats de la comparaison de l'efficacité des estimateurs HH et NS dans le cas d'enquêtes par échantillonnage à un degré et à deux degrés auprès des établissements au moyen du modèle d'erreur et expose brièvement les limites du modèle. La preuve d'un énoncé statistique figurant à la section 4.2 est donnée en annexe.

2. NOTATION

Représentons par N_j le nombre de ménages faisant des transactions avec l'établissement j ($j = 1, 2, \dots, R$), par N_o le nombre de ménages ne faisant pas de transactions avec un établissement et par N^* , le nombre de ménages distincts faisant des transactions avec R établissements. Alors, $N = N^* + N_o$ représente le nombre total de ménages. Représentons par M_j le nombre de transactions de l'établissement j ($j = 1, 2, \dots, R$) avec le ménage i , où $M_j \geq 0$ si l'établissement j fait des transactions avec le ménage i , et $M_j = 0$ si l'établissement j ne fait pas de transactions. Alors, $M_j = \sum_{i=1}^{N_j} M_{ij}$ représente le nombre de transactions de l'établissement j avec N ménages, et $M = \sum_{j=1}^R M_j$, le nombre de transactions de M établissements avec N ménages, et $M = M/N$, le nombre moyen de transactions par ménage. Représentons par X_k la valeur de la variable x pour la transaction k ($k = 1, \dots, M$) de l'établissement j ($j = 1, 2, \dots, R$). Alors, $X_j = \sum_{k=1}^{M_j} X_{jk}$ représente la somme de la variable x sur les M_j transactions de l'établissement j , et $X = \sum_{j=1}^R X_j$, la somme de la variable x sur les M transactions de R établissements. Représentons par $\bar{X}_j = X_j/M_j$ la valeur moyenne de la variable x sur les M_j transactions de l'établissement j , et $\bar{X} = X/M$, la valeur moyenne de la variable x sur M transactions.

3. ESTIMATEURS ET VARIANCES

3.1 Estimateur et variance HH

Considérons une enquête par échantillonnage à deux degrés autopondérée auprès des établissements réalisée à l'aide d'une liste autonome d'établissements qui énumère les R établissements et leur mesure de taille, M_j ($j = 1, 2, \dots, R$). Les établissements sont les unités primaires d'échantillonnage (UPP) et les transactions sont les unités secondaires d'échantillonnage. On sélectionne un échantillon PPT de r établissements avec remise à partir de la liste autonome, et on sélectionne indépendamment un

cherche à déterminer les catégories de fournisseurs de soins raisonnablement bonnes.

Ces dernières années, la recherche s'est concentrée sur les propriétés statistiques des estimateurs fondés sur des bases de sondage établies d'après des enquêtes démographiques et a pris une orientation plus théorique qu'apparavant. Les difficultés conceptuelles qu'a posé au départ l'élaboration d'estimateurs non biaisés pour une base de sondage établie d'après une enquête démographique, parce qu'un même établissement fait des transactions avec plusieurs ménages, ont été surmontées grâce à l'application de la théorie de l'échantillonnage en réseau (Sirken 1997; Thompson 1992). Sirken, Shimizu et Judkins (1995) ont élaboré la version par échantillonnage en réseau de l'estimateur HH, version que nous appellerons dans le présent article estimateur NS (pour *network sampling*), et Sirken et Shimizu (1999) ont élaboré la version par échantillonnage en réseau de l'estimateur d'Horwitz-Thompson (HT). Le présent article décrit le développement d'un modèle d'erreur statistique permettant de comparer l'efficacité de l'estimateur NS qui dépend de la base de sondage établie d'après une enquête démographique et celle de l'estimateur HH qui dépend de la base de sondage autonome, et de sa variance auprès des établissements fondée sur la base de sondage autonome, et de l'estimateur NS et de sa variance pour une enquête par échantillonnage à deux degrés auprès des établissements fondée sur la base de sondage établie d'après une enquête démographique.

La présentation de l'article est la suivante. La notation est décrite à la section 2. Les sections 3.1 et 3.2 donnent, respectivement, une description de l'estimateur HH autopondéré PPT et de sa variance pour une enquête par échantillonnage à deux degrés auprès des établissements fondée sur la base de sondage autonome, et de l'estimateur NS et de sa variance pour une enquête par échantillonnage fondée sur la base de sondage établie d'après une enquête démographique. L'élaboration du modèle d'erreur est présentée aux sections 4.1 à 4.4. La différence entre les variances HH et NS de deuxième degré pour des tailles d'échantillon prévues équivalentes est établie à la section 4.1. À la section 4.2, la composante de premier degré de la variance

Effets de plan de sondage dans les enquêtes auprès des établissements

MONROE G. SIRKEN¹

RÉSUMÉ

Lorsqu'on dispose de bases de sondage indépendantes énumérant tous les établissements et la mesure de leur taille, on utilise habituellement pour les enquêtes auprès des établissements l'estimateur PPT de Hansen-Hurwitz (HH) pour estimer le volume des transactions faites par les établissements avec les populations. Le présent article décrit la version par échantillonnage en réseau (NS pour *network sampling*) de l'estimateur HH proposée pour remplacer éventuellement la version PPT. L'estimateur NS dépend d'une liste d'établissements établie d'après une enquête démographique qui énumère les ménages et leur probabilité de sélection dans une enquête démographique par sondage et le nombre de transactions, si tant est qu'il y en ait, que fait chaque ménage avec chaque établissement. Un modèle statistique est élaboré en vue de comparer l'efficacité des estimateurs HH et NS en cas d'enquête par échantillonnage à un degré ou à deux degrés après des établissements en supposant que la base de sondage autonome et celle établie d'après une enquête démographique ne comportent aucune erreur de couverture ni de mesure de taille.

MOTS CLÉS : Bases de sondage autonome d'établissements; listes d'établissements établies d'après une enquête démographique; estimateur de Hansen-Hurwitz; estimateur par échantillonnage en réseau.

1. INTRODUCTION

Les listes d'établissements qui font des transactions avec des ménages visés par les enquêtes démographiques par sondage servent de bases de sondage aux enquêtes auprès des établissements lorsque l'on peut établir la correspondance entre les transactions déclarées par les ménages lors des enquêtes démographiques et les enregistrements des établissements pertinents. Par exemple, les listes d'établissements qui font des transactions avec les ménages participant à la National Medical Expenditure Panel Survey (MEPS), une enquête nationale auprès d'un échantillon de population, servent de bases de sondage pour les enquêtes sur les fournisseurs de soins médicaux qui permettent de compléter et de vérifier les données sur les dépenses pour soins médicaux correspondant aux transactions déclarées par les membres répondants des ménages de la MEPS (Cohen 1998). Cependant, les listes d'établissements qui font des transactions avec les ménages qui participent à des enquêtes démographiques par sondage servent rarement de bases de sondage aux enquêtes auprès des établissements destinées à recueillir des renseignements sur les transactions que les établissements font avec tous les ménages. Le Current Price Index (CPI) produit par le Bureau of Labor Statistics est un cas rare, qui mérite d'être mentionné, d'enquête auprès d'établissements fédéraux fondée sur une base de sondage produite à partir d'une enquête auprès d'un échantillon de population. La CPI Pricing Survey, une enquête nationale auprès des établissements de détail qui a pour but de recueillir des données sur les prix pour un panier de biens de consommation achetés par l'ensemble des consommateurs, a pour base de sondage les listes

d'établissements de détail qui font des transactions avec les ménages qui participent à la CPI Continuing Point of Purchase Survey (Leaver et Valliant 1995). Après avoir examiné les plans de restructuration de la famille d'enquêtes nationales indépendantes auprès des fournisseurs de soins de santé (hôpitaux, médecins, cliniques, etc.) du National Center for Health Statistics (NCHS), un groupe d'experts du Comité on National Statistics a proposé (Wunderlich 1992) d'utiliser les listes de fournisseurs de soins de santé déclarés par les ménages lors de la National Health Interview Survey (NHIS), qui est une enquête par sondage nationale permanente auprès des ménages (Massey, Moore, Parsons et Tadros 1991), comme bases de sondage des enquêtes nationales auprès des fournisseurs de soins de santé. Selon le Comité, étant donné l'évolution rapide des listes de fournisseurs de soins de santé due à l'évolution rapide du système national de prestation des services de santé, les listes de fournisseurs de soins de santé établies d'après la NHIS seraient plus exactes, et plus faciles et moins coûteuses à produire et à tenir à jour que les listes indépendantes de fournisseurs de soins de santé utilisées à l'époque. Peu après la diffusion du rapport du groupe d'experts, le NCHS a lancé un projet de recherche sur les bases de sondage produites d'après des enquêtes démographiques que nous résumons brièvement plus bas.

Au départ, l'étude s'est concentrée presque exclusivement sur les propriétés statistiques des listes de fournisseurs de soins de santé établies d'après la NHIS. Judkins, Berk, Edwards, Mohr, Stewart et Waksberg (1995) ont étudié la qualité des listes autonomes de fournisseurs de soins de santé utilisées à l'époque ou susceptibles d'être utilisées, et

¹ Monroe G. Sirken, Senior Research Scientist, National Center for Health Statistics, U.S.A.

- MCCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*. Deuxième Edition, London: Chapman and Hall.
- MURRAY, D. M., HANNAN, P. J., WOLFINGER, R. D., BAKER, W.L. et DWYER, J.H. (1998). Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine*, 17, 1581-1600.
- RUST, K.F., et RAO, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- SAS INSTITUTE INC. (1999). *SAS/STAT® User's Guide, Version 8*. Cary, NC: Author.
- SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SEARLE, S.R., CASELLA, G. et MCCULLOCH, C.E. (1992). *Variance Components*. New York: John Wiley & Sons Inc.
- SHAH, B.V., BARNWELL, B.G. et BIELER, G.S. (1997). *SUDAAN User' Manual, Release 7.5*. Research Triangle Park, NC: Research Triangle Institute.
- SHAH, B.V., HOLT, M. M. et FOLSOM, R.E. (1977). Inference About Regression Models from Survey Data. *Bulletin of the International Statistical Institute*, 41, 43-57.
- SHAPIRO, M.F., MORTON, S.C., MCCAFFREY, D.F., SENTERFIT, J.W., FLEISHMAN, J.A., PERLMAN, J.F., H. ATHEY, L.A., KEESY, J.W., GOLDMAN, D.P., BERRY, S. H. et BOZZETTE, S.A. (1999). Variations in the care of HIV-infected adults in the United States; results from the HIV cost and services utilization study. *Journal of the American Medical Association*, 281, 2305-2315.
- SKINNER, C.J. (1989a). Introduction to Part A. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, et T.M.F. Smith). New York: John Wiley & Sons Inc. 23-57.
- SKINNER, C.J. (1989b). Domain means, regression and multivariate analyses. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt et T.M.F. Smith). New York: John Wiley & Sons Inc. 59-88.
- STATACORP. (1999). *Stata Statistical Software: Release 6.0*. College Station, TX: Author.
- THEIL, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons Inc.
- WARE, J.E., JR., KOSINSKI, M. et KELLER, S.D. (1995). *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales*. Boston, Mass: The Health Institute, New England Medical Center.
- WELLS, K.B., SHERBOURNE, C., SCHOENBAUM, M., DUAN, N., MEREDITH, L., UNUTZER, J., MIRANDA, J., CARNEY, M. et RUBENSTEIN, L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association*, 283, 212-220.
- WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.
- WU, C.J.F., HOLT, D. and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-9.
- ZEGER, S.L., et LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.

Le terme correspondant à $r = 0$ est égal à $l'(\mathbf{X}'\mathbf{X})^{-1}l = \text{Var}(l' \beta)$. Le terme correspondant à $r = 1$ est nul. En vertu du théorème binomial,

$$\sum_{s=0}^{r+s} \binom{r}{r+s} \frac{1}{n^s} = \left(\frac{n}{n} \right)^{r+s},$$

de sorte que l'on peut apparier les autres termes, correspondant à $r = 2, 4, 6, \dots$, pour donner

$$\left(\frac{n}{n} \right)^r l' \left\{ \sum_{i=1}^r (\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i \right\} l^2 + \left(\frac{n-1}{n} \right)^{-1} \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1} \right] l^2 \left\{ \sum_{i=1}^r (\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1} \right\} l^2.$$

Le facteur médian de la somme peut s'écrire sous la forme

$$\left(\frac{n-1}{2} \right) (\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n-1}{n} \right) (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{X})^{-1},$$

qui est une définie positive, de sorte que l'expression entière doit être positive. Par conséquent, nous avons montré que $E(V_{jk}^2) \geq \text{Var}(l' \beta)$. L'égalité ayant lieu si, et uniquement si, $l'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X}$ est constante sur les i .

Preuve du théorème 4

$$\mathbf{v}^* = \mathbf{c}' \sum_{i=1}^l l' (\mathbf{X}\mathbf{X})^{-1} \mathbf{X}'_i \mathbf{A}_i (\mathbf{I} - \mathbf{H})^i \mathbf{e}' (\mathbf{I} - \mathbf{H})^i,$$

$$\mathbf{A}_i^t \mathbf{X}_i (\mathbf{X}\mathbf{X})^{-1} l'$$

$$= \mathbf{e}' \sum_{i=1}^l \mathbf{g}_i' \mathbf{g}_i' \mathbf{e}.$$

Soit \mathbf{P} , la matrice des vecteurs propres et \mathbf{A} , la matrice diagonale dont les éléments $\lambda_1, \dots, \lambda_M$ sont égaux aux valeurs propres de $\mathbf{V}^{1/2} \sum_{i=1}^l \mathbf{g}_i' \mathbf{g}_i' \mathbf{V}^{1/2} = \mathbf{B}' \mathbf{B}$ où $\mathbf{B}' = \mathbf{V}^{1/2} [\mathbf{g}_1' \mathbf{g}_2' \dots \mathbf{g}_l']$. Soit $\mathbf{u} = \mathbf{P}' \mathbf{V}^{-1/2} \mathbf{y}$ où $\mathbf{V}^{-1/2} \mathbf{V} \mathbf{V}^{-1/2} = \mathbf{I}$ définissant $\mathbf{V}^{-1/2}$; alors, les éléments de \mathbf{u} sont des variables normales indépendantes de variance 1 et

$$\mathbf{v}^* = \mathbf{u}' \mathbf{A} \mathbf{u} = \sum_{i=1}^l \lambda_i u_i^2.$$

Soit λ_i , toute valeur propre non nulle de $\mathbf{B}' \mathbf{B}$; alors, il existe un vecteur non nul \mathbf{z} tel que $\mathbf{B}' \mathbf{B} \mathbf{z} = \lambda_i \mathbf{z}$ et $\mathbf{B}' \mathbf{B} \mathbf{z} = \lambda_i \mathbf{B} \mathbf{z}$. Comme $\mathbf{B} \mathbf{z} \neq 0$, λ_i est une valeur propre de \mathbf{B}' . Par ailleurs, toute valeur propre non nulle de $\mathbf{B}' \mathbf{B}'$, $\mathbf{B}' \mathbf{B}$, sont égales aux valeurs propres non nulles de $\mathbf{B}' \mathbf{B}$. Par conséquent, les valeurs propres non nulles de $\mathbf{B}' \mathbf{B}$ sont égales aux valeurs propres non nulles de $\mathbf{B}' \mathbf{B}' = \{\mathbf{g}_i' \mathbf{V} \mathbf{g}_i\}$.

Association.

- MCCAFFREY, D.F., BELT, R.M. and BOTTS, C.H. (2001). Generalizations of bias reduced linearization. *Proceeding of the Survey Research Methods Section, American Statistical Association*.
- MCCAFFREY, D.F., BELT, R.M. (1997). Bias reduction in designs with few primary sampling units. Article présenté à Joint Statistical Meetings, Anaheim CA.
- MCCAFFREY, D.F., et BELT, R.M. (1997). Bias reduction in standard error estimates for regression analyses from multi-stage designs with improved small-sample properties. *Biometrics*, 57, 126-134.
- MANCL, L.A., et DEROUEN, T.A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 29, 305-325.
- MACKINNON, J.G., et WHITE, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305-325.
- 2, 92, Washington, D.C.: US Government Printing Office.
- STEHOUEW, S.A. (1982). A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and Health Statistics, Series*
- LANDIS, J.R., LEFKOWSKI, J.M., EKLAND, S.A. et LANDIS, J.R. (1982). A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and Health Statistics, Series*
- Société Statistique du Canada. 39-47.
- KOTT, P.S. (1996). Linear regression in the face of specification error: model-based exploration of randomization-based techniques. *Revue de la Section des méthodes d'enquêtes*, 20, 167-172.
- KOTT, P.S. (1994). Test d'hypothèse portant sur des coefficients de régression linéaire et basé sur des données d'enquête. *Techniques d'enquête*, 20, 167-172.
- KOTT, P.S. (1994). Test d'hypothèse portant sur des coefficients de régression linéaire et basé sur des données d'enquête. *Techniques d'enquête*, 20, 167-172.
- KORN, E.L., et GRAUBARD, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A, General*, 158, 263-295.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons Inc.
- GOLDSTEIN, H. (1991). Multilevel modeling of survey data. *The Statistician*, 40, 235-244.
- GELMAN, A., CARLIN, J.B., STERN, H.S., et RUBIN, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, 37, 117-32.
- ELLICKSON, P.L., et MCGUGAN, K.A. (2000). Early predictors of adolescent violence. *American Journal of Public Health*, 90, 566-572.
- ELLICKSON, P.L., et MCGUGAN, K.A. (2000). Early predictors of adolescent violence. *American Journal of Public Health*, 90, 566-572.
- EFRON, B., et TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- COOK, R.D., et WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Troisième édition, Park, CA: Sage.
- Models: Applications and Data Analysis Methods. Newberry Park, CA: Sage.
- BRYS, A.S., et RAUDEMENBUSH, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newberry Park, CA: Sage.
- TD-4S9HT, www.research.att.com/~rbel.
- Multi-Stage Samples. AT&T Labs-Research, Florham Park, NJ.
- Linearization Standard Errors for Linear Regression with Multi-Stage Samples. AT&T Labs-Research, Florham Park, NJ.
- BELT, R.M., et MCCAFFREY, D.F. (2002). Bias Reduction in Linearization Standard Errors for Linear Regression with Multi-Stage Samples. AT&T Labs-Research, Florham Park, NJ.

BIBLIOGRAPHIE

et Nelder 1989), le choix évident pour ce modèle consiste à utiliser la LBR fondée sur les poids finals et à fixer $\mathbf{U} = \mathbf{W}^{-1}$. Néanmoins, le théorème 3 ne se généralise pas aux MLG, parce que les poids sont estimés d'après les données et que nous n'avons pas étudié les propriétés de la LBR dans ce contexte.

Korn et Graubard (1995) proposent v_{12}^L comme estimateur de l'erreur-type pour les échantillons stratifiés dans des situations où la stratification est non-informative. Le même raisonnement s'applique à v_{12}^{LBR} . Fuller (1975) a proposé, pour les échantillons stratifiés, un autre estimateur de l'erreur-type convergent pour le plan d'échantillonnage. Bell et McCaffrey (2002, page 32 et 33) montrent que si l'on ajuste le vecteur des résidus de chaque strate, la LBR peut réduire ou éliminer le biais lié au modèle susceptible d'exister dans l'estimateur de Fuller.

REMERCIEMENTS

Nous remercions les examinateurs et le rédacteur adjoint de leurs commentaires constructifs au sujet d'une ébauche antérieure. Ces travaux sont financés en partie par la bourse 0001763 de la NSF.

ANNEXE

Preuves des théorèmes 2 et 4

Preuve du théorème 2. Suivant les premières étapes de la preuve du théorème 1, l'équation (6) implique que

$$E(v_{JK}) = \left(\frac{n}{n-1} \right)^L (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^I \mathbf{X}_i' \mathbf{d}_i' - \mathbf{H}^n \right)^{-1} \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1} L.$$

L'existence de $(\mathbf{I}^i = \mathbf{H}^n)^{-1}$ implique que les valeurs propres de \mathbf{H}^n sont strictement inférieures à 1, de sorte que $(\mathbf{I}^i = \mathbf{H}^n)^{-1}$ peut s'écrire sous la forme $\sum_{j=0}^{\infty} \mathbf{H}_j^n$. Conséquemment, en supposant que $\mathbf{D} = (1/n)$ et nous avons

$$\mathbf{D}^i = (\mathbf{X}_i' \mathbf{X}_i)^{-1} - \mathbf{D}, \text{ nous avons}$$

$$\left(\frac{n}{n-1} \right)^L (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^I \sum_{j=0}^{\infty} \mathbf{X}_i' \mathbf{X}_i - \mathbf{D} \right)^{-1} \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1} L =$$

$$\left(\frac{n}{n-1} \right)^L (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \mathbf{X}_i' \mathbf{X}_i - \mathbf{D} \right)^{-1} \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1} L =$$

$$\left(\frac{n}{n-1} \right)^L (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{s=0}^{\infty} \sum_{j=0}^{\infty} \mathbf{X}_i' \mathbf{X}_i - \mathbf{D} \right)^{-1} \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1} L =$$

X reste la même à mesure que n augmente (par exemple, par répétition), l'équation (4) implique que le biais diminue proportionnellement à $1/(n-1)$. En outre, les résultats observés pour $n = 20$ pourraient l'être pour une valeur beaucoup plus élevée de n si le gros de la variation de **X** est attribuable à quelques UPE et que la détermination de l/β dépend aussi d'un petit nombre d'UPE. Enfin, pour réduire le nombre de facteurs qui influent sur les résultats, nous avons simplifié le plan d'échantillonnage de plusieurs façons : taille constante des UPE, pas de coefficient de pondération ni de strates, et faible multicollinéarité. Nous avons le sentiment que relâcher l'une de ces contraintes aurait effectivement tendance à réduire la performance de la linéarisation standard et du jackknife. Selon nous, le choix de $m = 10$ pour la taille des UPE n , dans un sens ou dans l'autre, pas eu grand effet sur nos résultats.

Selon nous, les méthodes que nous proposons seront utiles aux analystes d'échantillons à plusieurs degrés, mais elles ne résoudront pas complètement le problème d'inférence dans le cas de la régression linéaire non pondérée. Nous avons tous deux observé fréquemment la situation perturbante où les méthodes de linéarisation standard produisent des intervalles de confiance plus petits que les méthodes qui ne tiennent pas compte du plan d'échantillonnage. Sans aucun doute, le biais de v_L et l'utilisation incorrecte de $n-1$ degrés de liberté contribuent à la fréquence de ce phénomène, mais nos méthodes ne l'empêchent pas de se manifester (voir la section 7). Comme les méthodes de réutilisation de l'échantillon, la linéarisation produit nécessairement des estimateurs dont la variance est grande pour certains, voire tous, les coefficients dans le cas de certains plans d'échantillonnage. Dans variable x_j , le nombre de degrés de liberté de Satterthwaite tombe presque à 3, ou devient inférieur à cette valeur, les analyses devraient se demander sérieusement s'ils peuvent accepter la grande variabilité, et la perte correspondante de puissance, liée à l'utilisation d'estimateurs non paramétriques de la variance. Des estimateurs paramétriques de régression, comme les modèles linéaires hiérarchiques ou l'inférence fondée sur l'estimation d'une corrélation intra-classes commune sur toutes les UPE (Wu et coll. 1988) devraient produire des résultats plus stables.

Le présent article porte sur la régression linéaire non pondérée pour des échantillons non stratifiés, mais nous n'avons aucune raison de penser que les problèmes du biais et du nombre de degrés de liberté que pose la linéarisation seraient atténués par la stratification, pour les moindres cartes pondérées ou pour des modèles linéaires généralisés (MLG). Comme l'ont montré McCaffrey, Bell et Bouts (2001), la méthode LBR se généralise immédiatement à la régression linéaire pondérée grâce à l'utilisation de $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ dans la condition principale du théorème 3. Puisque les solutions des MLG, comme la régression logistique, sont équivalentes aux étapes finales des moindres carrés itérativement pondérés (McCullagh

Par la méthode de Wu, Holt et Holmes (1988), nous calculons que la corrélation intra-clinique est égale à $-0,0026$, valeur qui concorde facilement avec une valeur réelle de 0. Néanmoins, il n'y a aucune raison de s'attendre à ce que toute erreur-type correcte soit nettement plus faible que celle obtenue par les MCO. La colonne 3 du tableau 4 montre que les erreurs-types produites par la méthode de linéarisation sont fréquemment très inférieures à celles obtenues par les MCO – particulièrement pour les variables indépendantes au niveau de l'UPB à la partie supérieure du tableau. Pareillement, la méthode de linéarisation en choisissant une distribution de référence t_{n-1}^* produit souvent des valeurs P beaucoup plus faibles que la méthode

des MCO.

La LBR donne de meilleurs résultats que la linéarisation standard. Les erreurs-types produites par la LBR sont systématiquement plus grandes et, parfois, considérablement plus grandes que les erreurs-types produites par la linéarisation. Par exemple, les estimations par LBR pour les variables indépendantes au niveau de l'UPB sont, en moyenne, supérieures de 15 % aux estimations obtenues par la méthode de linéarisation. Par ailleurs, les erreurs-types produites par la LBR pour les variables au niveau des UPB sont souvent plus faibles que les estimations par les MCO. Donc, même si, en principe, les estimateurs LBR sont quasiment sans biais, leur variabilité produit des estimations très faibles pour certains coefficients. La variabilité est également reflétée par le nombre de degrés de liberté qui est très faible pour les variables de bloc et, quoique plus grand pour les variables au niveau des patients, reste considérablement inférieur à 42, c'est-à-dire le nombre de grappes moins un. Le nombre de degrés de liberté est particulièrement petit, soit 7,6, pour la variable indépendante Noir (égal à un si le patient est un Afro-américain et à zéro autrement). Des tracés analogues à la figure 1 montrent que les valeurs de la variable Noir sont concentrées dans trois grappes. La variable Noir est égale à zéro pour tous les patients de 24 des 43 grappes, et 48 des 78 Afro-américains de l'échantillon se retrouvent dans trois grappes uniques. Comme nous en avons discuté aux sections 2 et 4, la concentration des Noirs dans un petit nombre de grappes donne lieu à une forte variance des deux estimateurs et à un biais important de l'estimation par linéarisation, situations que l'on peut toutes deux observer au tableau 4.

8. DISCUSSION

Bien que la linéarisation soit un bon outil qui produit des erreurs-types convergentes et des inférences valides à mesure qu'augmente le nombre d'UPB dans des échantillons à plusieurs degrés, ses utilisateurs devraient être conscients des problèmes qu'elle pose. Les variances estimées des coefficients de régression linéaire (y compris les moyennes par domaine) ont tendance à être biaisées par défaut – particulièrement si les coefficients (ou combinaisons linéaires de coefficients) dépendent en grande partie

de données provenant d'un petit nombre d'UPB. Selon le plan d'échantillonnage, un biais important peut persister, même quand le nombre total d'UPB est assez grand. La méthode du jackknife standard pour les échantillons à plusieurs degrés a tendance à produire un biais au moins aussi grand de signe opposé. Pareillement, l'utilisation d'une distribution t de référence dont le nombre de degrés de liberté est inférieur d'une unité au nombre d'UPB peut sous-estimer fortement l'incertitude associée à l'estimation de la variance. Comme les deux problèmes (biais et nombre surestimé de degrés de liberté) ont tendance à se produire simultanément dans le cas de la linéarisation, les intervalles de confiance et les tests statistiques fondés sur cette méthode pourraient être beaucoup trop libéraux.

La linéarisation à biais réduit (LBR) produit des estimations non biaisées de la variance lorsque les erreurs sont homoscedastiques et non corrélées, et a tendance à réduire fortement le biais pour d'autres structures de covariance étudiées dans le cadre de nos simulations. Dans nos simulations, la LBR produit systématiquement un biais inférieur de 90 % ou plus à celui associé à la linéarisation standard et a tendance à donner de nettement meilleurs résultats que la méthode de linéarisation corrigée proposée par Kott en 1994. Les résultats obtenus pour la LBR sont semblables à ceux obtenus pour la méthode de Kott de 1996.

Comparativement à la linéarisation standard, l'utilisation de la LBR et du nombre estimé de degrés de liberté de Satterthwaite améliore considérablement l'inférence statistique. La réduction du biais et l'utilisation du nombre de degrés de liberté de Satterthwaite semblent contribuer de façon égale à l'amélioration des résultats. Bien que l'approximation de Satterthwaite puisse produire une surcom-pensation, menant à une inférence prudente dans certaines situations, le problème ne semble pas important tant que le nombre de degrés de liberté de Satterthwaite ne devient pas inférieur à cinq (fondé, en partie, sur des simulations non présentées ici). Le cas échéant, les analystes pourraient choisir d'estimer les valeurs critiques au moyen de simulations fondées sur le théorème 4.

Il est important de souligner certaines limites de nos résultats de simulations. En premier lieu, nous ne présentons les résultats que pour quatre variables indépendantes distinctes et une ordonnée à l'origine. Nous choisissons ces variables de façon à couvrir une grande gamme de situations. Bien que certains analystes puissent considérer x_2 comme extrême ou anormal, cette variable ne sort pas de la gamme de situations que nous avons observées dans le cadre de nos propres travaux de consultation. Des variables comme x_2 peuvent résulter d'essais randomisés par groupe (voir la section 7), de données d'observation lorsque quelques UPB seulement présentent un intérêt particulier ou de l'utilisation d'une série de variables nominales pour représenter les niveaux d'une variable catégorique. En deuxième lieu, nous présentons les résultats uniquement pour $n = 20$ UPB. Dans la mesure où

mois de suivi. Les scores ont été normalisés de sorte que la population générale, la santé étant d'autant meilleure que le score est élevé. Comme dans Wells et coll. (2000), la variable indépendante présentant le plus grand intérêt est un indicateur d'intervention qui estime l'effet combiné du traitement médicamenteux et de la thérapie par opposition aux soins habituels. Les deux premières colonnes du tableau 4 montrent les coefficients MCO et les erreurs-types pour l'effet d'intervention et toutes les covariables utilisées, mais non présentées, par Wells et coll. (2000). Notre régression diffère de la leur parce que nous ne corrigeons pas la non-réponse par pondération et n'imputons pas de données pour remplacer les valeurs manquantes de la

variable étudiée, mais les résultats pour l'effet d'intervention concordent raisonnablement. Comme les patients provenant d'une même clinique pourraient présenter des résultats comparables, les erreurs-types par les MCO pourraient facilement être trop faibles – particulièrement pour les variables au niveau de l'UPF, comme l'intervention. Les colonnes 3 et 4 du tableau 4 montrent les ratios des erreurs-types pour la linéarisation et la LBR aux erreurs-types pour les MCO. Nous choisissons la clinique comme UPF, parce qu'il y a fort peu de raisons de s'attendre à une corrélation des erreurs entre cliniques après que l'on ait tenu compte de l'effet des blocs.

Tableau 4

Comparaison de l'inférence par les MCO, la linéarisation et la LBR pour l'exemple de Partner-in-Care

Variable indépendante	β_j	$E.-T.^{MCO}$	$\frac{E.-T.^{LIN}}{E.-T.^{MCO}}$	$\frac{E.-T.^{LBR}}{E.-T.^{MCO}}$	DL _{LBR}	MCO	LIN	LBR
Niveau de l'UPF	28,795	3,409	1,03	1,06	23,7	0,000	0,000	0,000
Ordonnée à l'origine	1,724	0,746	0,73	0,84	15,4	0,021	0,003	0,015
Bloc 1	1,386	1,867	0,63	0,80	2,7	0,458	0,244	0,426
Bloc 2	-0,031	1,576	0,88	1,07	3,6	0,984	0,982	0,986
Bloc 3	-1,042	1,230	0,53	0,61	3,9	0,397	0,117	0,241
Bloc 4	0,038	1,231	0,62	0,73	4,5	0,976	0,961	0,968
Bloc 5	-3,707	1,503	0,66	0,78	4,7	0,014	0,001	0,027
Bloc 6	-0,025	1,562	1,15	1,32	4,9	0,987	0,989	0,991
Bloc 7	-2,784	1,644	0,84	0,97	7,0	0,090	0,051	0,126
Bloc 8	0,822	1,233	0,93	1,03	12,0	0,505	0,476	0,527
Démographique	0,972	1,448	0,74	0,79	7,6	0,502	0,369	0,419
Noir(e)	0,202	1,004	0,73	0,75	24,3	0,841	0,785	0,791
Autre non-Blanc(he)	-1,033	1,409	0,77	0,80	21,6	0,463	0,349	0,369
Femme	-0,502	0,803	1,09	1,12	23,1	0,532	0,571	0,581
Log de l'avoir net + 1 000\$	0,015	0,215	0,87	0,89	23,6	0,943	0,936	0,937
Pas de diplôme d'études secondaires	-1,690	1,217	1,00	1,04	25,3	0,165	0,173	0,192
Certaines études collégiales	-1,140	0,879	0,77	0,78	26,0	0,195	0,097	0,108
Diplôme collégial	-0,703	1,047	0,78	0,79	21,1	0,502	0,393	0,404
Âge	0,059	0,032	0,91	0,93	26,5	0,064	0,047	0,056
Mariée(e)	0,541	0,748	1,05	1,07	28,5	0,470	0,496	0,504
Indicateurs de base de la santé	-0,973	1,039	0,92	0,94	23,7	0,349	0,313	0,327
1 problème de santé chronique (sur 19)	0,198	1,116	0,87	0,90	23,0	0,859	0,84	0,846
2 problèmes de santé chroniques	-0,201	1,132	0,90	0,91	24,0	0,859	0,844	0,847
3 problèmes de santé chroniques ou plus	-5,305	1,335	0,93	0,95	25,8	0,000	0,000	0,000
Dépression et dysthymie	-3,882	0,982	1,12	1,15	23,7	0,000	0,001	0,002
Dépression et dysthymie	-2,396	1,109	1,02	1,05	21,2	0,031	0,04	0,052
Composante mentale du SF-12	0,287	0,036	1,11	1,14	26,6	0,000	0,000	0,000
Composante physique du SF-12	0,079	0,036	0,88	0,89	24,6	0,029	0,017	0,022
Trouble anxieux	-2,438	0,749	1,20	1,23	26,3	0,001	0,010	0,014

contre, la probabilité d'une erreur de première espèce est un plus petit nombre de degrés de liberté combiné à un biais positif important produit des valeurs de test très prudentes.

Taux d'erreurs de première espèce pour les tests de vérification de l'hypothèse nulle $\beta = 0$

Estimateur	DL	β_0	β_1	β_2	β_3	β_4
------------	----	-----------	-----------	-----------	-----------	-----------

$p = 0$

Linéarisation	n-1	7,54	7,00	15,99	7,35	5,38
Sat		5,75	6,45	6,33	6,28	5,18
Jackknife	n-1	5,01	3,92	7,58	4,52	5,02
Jackknife	Sat	3,80	3,43	1,41	3,26	4,77
Kott (1994)	Kott	4,87	5,03	7,13	5,21	4,67
LBR	n-1	6,28	5,37	11,25	5,90	5,21
LBR	Sat	4,73	4,86	3,12	4,72	5,00

$p = 1/9$

Linéarisation	n-1	7,81	7,14	16,19	8,18	5,34
Sat		6,03	6,60	6,43	7,05	5,14
Jackknife	n-1	5,31	4,06	7,63	4,49	4,77
Jackknife	Sat	4,11	3,61	1,48	3,24	4,51
Kott (1994)	Kott	5,07	5,03	7,00	5,51	4,56
LBR	n-1	6,52	5,50	11,27	6,23	5,08
LBR	Sat	5,04	5,00	3,19	4,93	4,84

$p = 1/3$

Linéarisation	n-1	8,10	7,28	16,39	8,79	5,66
Sat		6,30	6,78	6,62	7,53	5,44
Jackknife	n-1	5,45	4,11	7,76	4,56	4,67
Jackknife	Sat	4,13	3,61	1,51	3,35	4,46
Kott (1994)	Kott	5,14	5,06	7,02	5,80	4,84
LBR	n-1	6,76	5,63	11,55	6,45	5,19
LBR	Sat	5,18	5,14	3,30	5,26	4,98

Nota: Les entrées dont la valeur réelle est de 5,00 % ont une erreur-type de 0,07 %.

La LBR avec $(n - 1)$ degrés de liberté donne de meilleurs résultats que la linéarisation avec le même nombre de degrés de liberté. Comme la LBR est sans biais quand $p = 0$, la comparaison de la cinquième à la première ligne du tableau montre la réduction du taux d'erreurs de première espèce qui résulte de l'élimination du biais de linéarisation. Sauf pour β_3 , la LBR réduit le taux d'erreurs de première espèce de 45 % à 88 %. Cependant, la LBR avec $(n - 1)$ degrés de liberté continue de produire des résultats systématiquement libéraux, particulièrement pour β_2 . La comparaison des lignes 2 et 5 de chaque section du tableau montre l'effet relatif de la réduction du biais et

de l'ajustement de Satterthwaite. Pour β_0 et β_2 , le nombre de degrés de liberté est le facteur le plus important, tandis que pour β_1 et β_3 , c'est le biais qui importe. Pour la LBR avec l'approximation de Satterthwaite, les résultats sont très bons, sauf pour β_2 , pour lequel le taux d'erreurs de première espèce tombe à environ 3 %.

Les tests fondés sur l'estimateur de Kott de 1994 avec le nombre de degrés de liberté proposé donnent de très bons résultats pour les coefficients dont l'estimateur de la variance est biaisé par excès. Il semble que le biais par excès de l'estimateur de la variance soit compensé par le biais par excès du nombre approximatif de degrés de liberté. L'estimateur de la variance de Kott présente un léger biais négatif pour β_2 et, donc, le biais par excès du nombre de degrés de liberté aggrave le biais de l'estimateur pour aboutir à un taux d'erreurs de première espèce d'environ 7 % pour les trois valeurs de p .

Les tests basés sur l'estimateur de Kott de 1996 donnent aussi de bons résultats. Pour presque tous les coefficients et toutes les valeurs de p , le taux d'erreurs de première espèce s'approche de 5 %. Fait exception le test pour β_3 , lorsque $p = 1/3$, pour lequel le taux d'erreurs est de 5,88 % à cause du biais modéré de l'estimateur de la variance.

7. EXEMPLE TIRÉ DE L'EXPÉRIENCE PARTNERS IN CARE

Nous illustrons les méthodes décrites dans l'article à l'aide de données provenant de Partners in Care, une expérience longitudinale réalisée en vue d'évaluer l'effet des programmes d'amélioration de la qualité des soins prodigués aux déprimés par les organismes de gestion intégrée des soins de santé (OGIFS) (Wells, Sherbourne, Schoenbaum, Duan, Meredith, Unutzer, Miranda, Camery, Rubenstein 2000). L'expérience consistait à suivre 1 356 patients provenant de 43 cliniques de sept OGIFS chez lesquels le dépistage de la dépression avait été positif en 1996-1997. Les cliniques ont été réparties au hasard entre trois cellules expérimentales : soins habituels, programme d'amélioration de la qualité complète par des ressources pour le suivi du traitement médicalement et programme d'amélioration de la qualité complète par des ressources pour l'accès à la psychothérapie. Cette répartition a eu lieu après formation de 27 ensembles de cliniques – trois pour chacun de neuf blocs (six OGIFS représentèrent, chacun, un bloc unique et un OGIFS a été scindé en trois blocs d'après la composition ethnique des cliniques). Dans les blocs comptant plus de trois cliniques, les ensembles de cliniques ont été combinés de façon à obtenir la meilleure concordance possible avec la taille prévue d'échantillon et les caractéristiques des patients. Voir Wells et coll. (2000).

Nous présentons les résultats d'une régression par les MCO sur le score sommaire de santé mentale du SF-12 (Ware, Kosinski et Keller 1995) pour 1 048 patients à six

La LBR donne uniformément de meilleurs résultats, le pire biais qu'elle produit étant de -2,1 %. Alors que Kott (1996) ne diffère pour ainsi dire pas de la LBR pour les variables au niveau des UPE, il donne des résultats nettement moins bons pour β_3 et β_4 .

Tableau 1

Biais des estimateurs de la variance (en pourcentage de la variance réelle)	
Estimateur	
β_0	β_1
β_2	β_3
β_4	

MCO	0,0	0,0	0,0	0,0	0,0
Linéarisation	-9,6	-13,2	-32,5	-13,3	-1,8
Jackknife	11,7	17,2	51,2	17,6	2,1
Kott (1994)	4,0	2,5	-10	2,2	4,7
(erreurs-type)	(0,2)	(0,1)	-3	(0,2)	(0,1)
LBR	0,0	0,0	0,0	0,0	0,0
Kott (1996)	0,0	0,0	0,0	0,0	0,0
(erreurs-type)	(0,2)	(0,1)	-3	(0,2)	(0,1)

MCO	-50,2	-49,7	-50,7	-37,7	4,1
Linéarisation	-10,3	-14,2	-33,2	-17,1	-2,5
Jackknife	11,0	16,4	50,1	19,8	3,2
Kott (1994)	3,9	2,7	-0,8	1,5	4,6
(erreurs-type)	(0,2)	(0,1)	(0,3)	(0,2)	(0,1)
LBR	-0,7	-1,0	-0,8	-1,2	0,1
Kott (1996)	-0,8	-1,2	-1,0	-4,4	-0,7
(erreurs-type)	(0,2)	(0,1)	(0,3)	(0,2)	(0,1)

MCO	-75,8	-75,5	-76,2	-65,3	13,8
Linéarisation	-10,7	-14,8	-33,5	-19,9	-4,1
Jackknife	10,7	15,9	49,5	21,4	5,9
Kott (1994)	3,6	2,4	-0,6	1,4	4,4
(erreurs-type)	(0,2)	(0,1)	(0,3)	(0,2)	(0,1)
LBR	-1,2	-1,9	-1,5	-7,7	-2,3
Kott (1996)	-1,0	-1,5	-1,3	-2,1	0,4
(erreurs-type)	(0,2)	(0,1)	(0,3)	(0,2)	(0,1)

Nota: Toutes les valeurs sont exactes sauf pour Kott (1994), qui est fondé sur 100 000 répétitions de la simulation.

Les estimateurs par linéarisation, par le jackknife, par LBR et de Kott sont fortement corrélés et présentent des coefficients de variation similaires. Pour tout coefficient de régression, la corrélation entre les estimateurs de la variance excède systématiquement 0,969, la plupart des valeurs étant supérieures à 0,99 (données non présentées). Les corrélations les plus faibles ont tendance à être celles observées entre le jackknife et d'autres estimateurs. Les coefficients de variation (données également non présentées) les plus importants sont ceux obtenus pour Kott (1994) et ont tendance à être les plus faibles pour la linéarisation et Kott (1996) (sauf pour l'ordonnée à l'origine). Pour l'ordonnée à l'origine, le jackknife est l'estimateur qui donne le coefficient de variation le plus faible. La variance relative de l'estimateur par la LBR est comparable à celle des méthodes non paramétriques de recharge. Son coefficient de variation excède de 1 % à 6 % celui de l'estimateur par linéarisation, mais est de 5 % à 10 % plus faible que celui de l'estimateur Kott (1994). Donc, les cinq estimateurs non

paramétriques de la variance ont tendance à différer l'un de l'autre principalement par des facteurs constants. Le tableau 1 résume les principales différences entre ces estimateurs de la variance.

Le tableau 2 montre le nombre de degrés de liberté de Satterthwaite pour chacun des cinq coefficients pour les estimateurs de la variance par linéarisation, par le jackknife, par LBR et de Kott. Pour tous les estimateurs, nous avons calculé le nombre de degrés de liberté en supposant que $V = I$ et, conséquemment, il dépend uniquement de la matrice de plan d'échantillonnage et non des valeurs de y . Les approximations sont semblables pour la linéarisation et la LBR, quoique le nombre de degrés de liberté pour la linéarisation ait tendance à être un peu plus grand, reflétant le fait que, pour la matrice de plan d'échantillonnage considérée, les variances relatives des estimateurs LBR sont légèrement plus grandes que pour les estimateurs par linéarisation. L'approximation de Kott consiste à calculer le coefficient de variation pour un estimateur de type linéarisation d'après les erreurs réelles plutôt que les résidus. Par conséquent, le nombre approximatif de degrés de liberté de Kott, qui est plus grand que ceux obtenus pour la linéarisation ou la LBR, a tendance à surestimer la précision de cet estimateur (voir Kott 1994, section 6). Pour chacun des quatre estimateurs, les approximations les plus faibles sont celles obtenues pour β_2 .

Tableau 2

Nombre de degrés de liberté de certains estimateurs					
Méthode	β_0	β_1	β_2	β_3	β_4
Satterthwaite (LIN)	9,02	14,45	3,3	11,56	16,65
Satterthwaite (Jackknife)	9,52	13,3	2,62	9,06	16,23
Satterthwaite (LBR)	9,24	14,08	2,9	10,26	16,45
Méthode de Kott	10,33	16,41	4,32	11,36	17,44

Le tableau 3 montre que le taux d'erreurs de première espèce pour la méthode de linéarisation standard avec $(n - 1)$ degrés de liberté excède systématiquement 5 % pour les trois valeurs de p . C'est pour β_2 que les erreurs de première espèce sont les plus courantes, leur taux pouvant atteindre 16 %, mais elles sont également beaucoup trop fréquentes pour β_0 , β_1 et β_3 , leur taux variant de 7,0 % à 8,8 %. L'importance de ce problème est fortement corrigée à la suite du biais de l'estimateur par linéarisation (voir le tableau 1). Le taux d'erreurs de première espèce est nettement plus faible, de 5,7 % à 6,4 %, pour les tests où le nombre de degrés de liberté correspond à l'approximation de Satterthwaite. Donc, l'utilisation de l'autre nombre de degrés de liberté améliore de 30 % à 88 % le taux d'erreurs de première espèce.

Le profil des probabilités d'erreur de première espèce est moins uniforme pour la méthode du jackknife. Avec $(n - 1)$ degrés de liberté, cette méthode a tendance à produire, pour β_1 et β_3 , des résultats prudents en harmonie avec le biais positif de l'estimation de la variance par le jackknife. Par

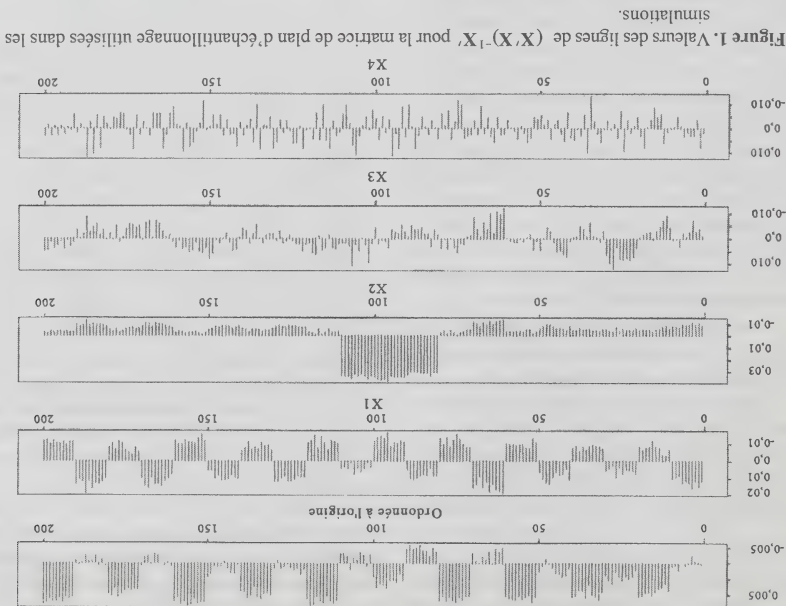


Figure 1. Valeurs des lignes de $(X'X)^{-1}X'$ pour la matrice de plan d'échantillonnage utilisées dans les simulations.

Nous estimons le taux d'erreurs de première espèce pour huit variantes, basées sur 100 000 répétitions, du test de vérification de l'hypothèse nulle $\beta_k = 0$, pour $k = 0$ à 4. Chaque variante consiste à comparer une « statistique t » à une distribution t de référence. Pour les statistiques fondées sur la linéarisation, le jackknife et la LBR, nous utilisons les valeurs critiques des distributions t pour un nombre de degrés de liberté égal à $(n - 1) = 19$, ainsi qu'à l'approximation de Satterthwaite correspondante. Pour les méthodes de Kott, nous utilisons le nombre de degrés de liberté proposé par ce dernier. Tous les calculs ont été réalisés en SAS.

6. RÉSULTATS DES SIMULATIONS

Le tableau 1 montre le biais de plusieurs estimateurs de la variance pour les cinq coefficients de régression (γ compris l'ordonnée à l'origine) pour $p = 0,1/9$ et $1/3$. Sauf pour Kott (1994), toutes les valeurs sont exactes d'après la matrice X décrite plus haut. Comme Kott (1994) ne peut pas s'exprimer sous forme de fonction linéaire, son biais est estimé d'après les simulations de Monte Carlo et l'erreur-type de ce biais est indiquée entre parenthèses. Les variances MCO sont sans biais pour $p = 0$, mais fortement biaisées pour $p = 1/9$ et $1/3$. Comme en discutent Wu, Holt et Holmes (1988), les variances MCO sont sous-estimées d'environ un facteur $1/[1 + p(m - 1)ICC]$, où ICC_x représente la corrélation intra-grappe d'une

Par conception, Kott (1996) et la LBR éliminent le biais l'estimateur de Kott demeure biaisé par défaut. Fait exception β_2 , pour lequel un biais positif important. méthode à surcompenser, ce qui produit souvent faible que celui produit par la linéarisation. Cependant, la méthode à tendance à surcompenser, ce qui produit souvent un biais positif important. Fait exception β_2 , pour lequel l'estimateur de Kott demeure biaisé par défaut.

Sauf pour β_4 , le biais de Kott (1994) est nettement plus observé pour β_3 , suivi de près par ceux de β_1 et β_0 . trois UPE. Vient ensuite, par ordre décroissant, le biais la valeur dépend principalement des données provenant de importants (en valeur absolue) sont observés pour β_2 , dont fortement d'un coefficient à l'autre. Les biais les plus Pour chaque estimateur, la grandeur du biais varie relativement indépendants de p , mais de signes opposés. présentent tout deux un biais important; ces biais sont Les estimateurs par linéarisation et par le jackknife négative pour x_4 .

de β_4 résulte de la corrélation intra-grappe légèrement la variable de niveau individuel pour laquelle la corrélation intra-grappe est grande. Le biais positif de la variance MCO façon, le biais est plus faible, mais reste considérable pour x_3 , sous-estimées d'environ un facteur 1/DEFF. De la même compris l'ordonnée à l'origine), les variances MCO sont variable x . Donc, pour les variables au niveau de l'UPE (γ

la forme

singulière V . Alors, pour tout estimateur de la variance de

$$v^* = c'l'(X'X)^{-1}\sum_{i=1}^l X_i' A_i r_i r_i' A_i' X_i' (X'X)^{-1} l_i,$$

v^* est égal à la somme pondérée de variables aléatoires indépendantes X_i' où les poids sont les valeurs propres de la matrice $n \times n$ $G = \{g_i' A_i g_i\}$, pour $g_i = c^{1/2} (I - H)^i A_i X_i' (X'X)^{-1} l_i$ (preuve en annexe).

Nous pouvons écrire v_L sous la forme quadratique $y' G^* y$, où la matrice M -par- M $G^* = \sum_{i=1}^n g_i g_i'$, de sorte que v_L est une somme pondérée de variables aléatoires khi-deux indépendantes dont les poids sont égaux aux valeurs propres de $G^* V$. La preuve consiste à montrer que les valeurs propres non nulles de $G^* V$ sont égales aux valeurs propres non nulles de G .

La moyenne et la variance de v^* sont de simples fonctions des valeurs propres de G , à savoir $E(v^*) = \sum_{i=1}^n \lambda_i^2 E(u_i^2) = \sum_{i=1}^n \lambda_i^2 \text{Var}(v^*) = \sum_{i=1}^n \lambda_i^2 \text{Var}(u_i^2)$. Si $V = \sigma^2 I$ et $X_i' X_i' (X'X)^{-1} l_i$ pour $i = 1, \dots, n$ sont constantes, condition nécessaire pour que v_L et v_{JK} soient sans biais, alors la théorie 4 implique que $a v_L^2, a v_{JK}^2$, et $a v_{LBR}$ suivent tous la loi de distribution χ_{n-1}^2 pour $a = (n-1)/\text{Var}(\beta)$ (Bell et McCaffrey 2002, page 41 et 42). Cependant, en général, $X_i' X_i' (X'X)^{-1} l_i$ n'est pas constante et le carré du coefficient de variation est supérieur à $2/(n-1)$, la statistique correspondante pour la variable aléatoire χ_{n-1}^2 .

Il convient notamment de se préoccuper de cette variabilité excédentaire lorsqu'on envisage, pour le test de vérification de l'hypothèse nulle $\beta = 0$, des distributions de cation de l'essai. Holt et Folsom (1977) proposent de comparer t à une distribution t de référence à $n-1$ degrés de liberté, qui est maintenant la distribution par défaut dans Stata (Stata Corp. 1999), dans SUDAAN (Shah, Barnwell et Bieler 1997) et dans SAS (SAS Institute 1999). Le choix de $n-1$ degrés de liberté est motivé par le fait que v_L peut s'écrire sous forme de somme des carrés de n variables aléatoires $c^{1/2} l_i' (X'X)^{-1} X_i' r_i r_i' X_i' (X'X)^{-1} l_i$. Cependant, comme la variance de $(n-1) v_L / E(v_L)$ a tendance à être plus grande que $2(n-1)$, les tests basés sur une distribution t à $n-1$ degrés de liberté auront tendance à produire un taux d'erreurs de première espèce supérieur à la valeur nominale, même si v_L est sans biais.

Satterthwaite (1946) a proposé d'approximer la distribution d'une combinaison linéaire de variables X_i' par $\chi_{f'}^2$ (jusqu'à une constante) où les deux premiers moments de la combinaison linéaire concordent avec ceux de $\chi_{f'}^2$. Nous approchons ainsi v_L , v_{LBR} ou v_{JK} par une $\chi_{f'}^2$ où $f = 2/cv^2 = (\sum_{i=1}^n \lambda_i^2) / (\sum_{i=1}^n \lambda_i^2)$ et où les valeurs propres de la matrice G correspondent. Les tests basés sur les distributions t de référence ayant f degrés de liberté devraient, en principe, donner un meilleur taux d'erreurs de première espèce que ceux fondés sur $n-1$ degrés de liberté. Rust et Rao (1996) proposent aussi d'utiliser une approximation de Satterthwaite pour estimer le nombre de degrés

choix de A_i est sans importance, car toute solution de (7)

produit un estimateur sans biais de la variance. Cependant, les estimateurs résultants sont biaisés lorsque $V \neq \sigma^2 I$, et le biais peut varier fortement selon le choix de A_i . Heureusement, il est sensé de choisir la solution A_i « la plus proche » de la matrice d'identité, afin de « mélanger » les

résidus aussi peu que possible. Deux candidats prometteurs sont la décomposition de Cholesky de $(I - H)^{-1}$, où tous les 0 se trouvent sous la diagonale, et la racine carrée symétrique de $(I - H)^{-1}$. Soit P , une matrice orthogonale dont les colonnes sont les vecteurs propres de $(I - H)^{-1}$ et A , une matrice diagonale contenant les valeurs propres correspondantes de $(I - H)^{-1}$, de sorte que $(I - H)^{-1} = P A P'$. Alors, pour $A^{1/2}$ égale à la racine carrée de A en ce qui concerne les éléments, $P A^{1/2} P'$ est symétrique et résout (7). Par contre, la multiplication de l'une ou l'autre de ces deux solutions par une matrice diagonale aléatoire pourrait fausser considérablement les résidus.

Parmi la classe de résidus ajustés de la forme $A_i r_i r_i'$ où A_i satisfait (7), ceux basés sur la racine carrée symétrique de $(I - H)^{-1}$, $r_i' = P A^{1/2} P' r_i$, sont « les meilleurs » au sens de Theil (1971) – c'est-à-dire qu'ils minimisent la somme attendue des carrés des différences entre les erreurs estimées et les erreurs réelles i.i.d. (pour les détails, voir page 36 et 37 dans Bell et McCaffrey 2002). En cas de corrélation intra-grappe, les résultats de simulations de la section 6 donnent à penser que le biais de v_L , fondé sur la racine carrée symétrique est nettement plus faible que celui de l'estimateur par linéarisation standard, v_L . Par conséquent, nous considérons uniquement la racine symétrique de l'article et nommons l'estimateur obtenu en utilisant cette racine l'estimateur par linéarisation à biais

réduit, v_{LBR} . Comme Kott (1994) l'a prouvé pour v_L , si le nombre d'unités dans chaque UPE est borné et que les éléments de $(X'X)^{-1} X_i' r_i r_i' X_i' (X'X)^{-1}$ sont bornés par B/n pour une constante donnée B (c'est-à-dire, $(X'X)^{-1} X_i' r_i r_i' X_i' (X'X)^{-1}$ est $O(1/n)$), alors le biais de v_{LBR} est $O(n^{-2})$ et le biais relatif est $O(1/n)$ (Bell et McCaffrey 2002, page 15).

4. VARIANCE DES ESTIMATEURS ET TESTS

Nous notons que v_L, v_{LBR} , et v_{JK} peuvent tous s'écrire sous la forme
$$v^* = c'l'(X'X)^{-1}\sum_{i=1}^l X_i' A_i r_i r_i' A_i' X_i' (X'X)^{-1} l_i,$$
 où $c = n/(n-1)$, 1 ou $(n-1)/n$, respectivement, et $A_i = I^{1/2} (I - H^{11})^{-1/2} (I - H^{11})^{-1}$ ou $(I - H^{11})^{-1}$, respectivement. Cette formulation des estimateurs montre que v_{LBR} peut être considéré comme un compromis entre v_L et v_{JK} choisi pour compenser les biais opposés de ces derniers.

Théorème 4

Soit les termes d'erreur distribués selon une loi normale multivariée de moyenne 0 et de matrice de covariances non

L'exemple montre que, si les erreurs sont i.i.d., v_L n'est

sans biais que dans des conditions très contraignantes.

Quand $V \neq I$, les théorèmes 1 et 2 ne sont pas vérifiés et le

biais de v_L peut même être positif (voir l'exemple 2 de Bell

et McCaffrey 2002).

En général, v_L est biaisé négativement. L'estimateur est

égal à la somme sur les UPB des carrés des combinaisons

linéaires des résidus, $c^{1/2} I' (X' X)^{-1} X' r$. Ces sommes des

carrés ont tendance à être trop faibles pour deux raisons :

les résidus sont généralement plus petits que les erreurs

réelles à cause du surajustement du modèle, et leur corré-

lation intra-grappe à tendance à être plus faible que celle

des erreurs. Le facteur $c = n/(n-1)$ ne corrige entièrement

ces problèmes que dans des circonstances fort restreintes,

comme les conditions du théorème 1.

Le biais de l'estimateur par linéarisation (ou par le

jackknife) augmente parallèlement à la variance inter-UPB

des variables indépendantes. Par conséquent, les variables

indépendantes dont la valeur dans les UPB est (presque)

constante ont tendance à présenter le biais le plus important.

Si plusieurs variables indépendantes de ce type existent, la

sous-estimation des corrélations intra-grappe peut être con-

siderable et biaiser fortement les variances estimées pour

tous les coefficients correspondants. Un risque encore plus

grand de biais semble exister lorsque certaines UPB sont la

cause de la plupart de la variabilité des covariables et

qu'elles ont un effet disproportionné sur la détermination de

3. LA MÉTHODE DE LINÉARISATION À BIAIS RÉDUIT

Phillip Kott a proposé deux méthodes pour réduire le

biais de v_L en utilisant les résidus et la matrice de plan

d'échantillonnage pour estimer la négative du biais de v_L

par $R(R > 0)$, habituellement) et en fixant $v_{K94} = v_L$

(1 - R/v_L). Il propose l'estimateur v_{K94} plutôt que l'estima-

teur plus évident ($v_L + R$) comme correction ponctuelle du

biais relatif de R en tant qu'estimateur du biais négatif réel,

Dans son article de 1996, Kott propose de calculer le

ratio de Var(I'/β) à $E(v_L)$ sous l'hypothèse que $V = I$ et de

rajouter v_L par le ratio. Si $V = I$, alors l'estimateur résultant

v_{K96} sera sans biais.

Dans le contexte des équations d'estimation généra-

lisées, Mancl et DeKouen (2001) adoptent une méthode

différente pour corriger le biais de l'estimateur par

linéarisation. Ils proposent d'ajuster les résidus provenant

est égal à $n/(n-1)^{1/2} I$ et ses propriétés découlent de celles

de l'estimateur par le jackknife.

Nous présentons une autre méthode, que nous avons

proposée pour la première fois en 1997 (McCaffrey et Bell

1997). Elle s'appuie aussi sur le remplacement de r dans

l'équation (2) par des résidus ajustés de la forme $r_i^* = A_i r_i$

destinés à agir davantage comme les erreurs réelles ε_i . À

l'instar de Kott (1996), nous calculons un estimateur qui

élimine le biais de v_L lorsque V égale U , une matrice dia-

gonale par blocs des covariances, et qui le réduit pour les

autres V . Comme Mancl et DeKouen (2001), nous ajustons

les résidus pour chaque UPB. Cependant, partant de U ,

nous calculons une autre approximation de $E(r_i r_i')$ et, au

lieu d'être proportionnel au jackknife, notre estimateur peut

être considéré comme un compromis entre les estimateurs

par linéarisation et par le jackknife. Notre méthode est

également une généralisation de la méthode de MacKinnon

et White (1985), qui ajustent les résidus individuels pour

produire un estimateur de la variance hétéroscédastique-

ment convergent (au sens de White 1980) dépourvu de biais

lorsque les erreurs sont indépendantes et homoscédastiques.

Théorème 3

Pour une matrice diagonale par blocs des covariances

précisée U , considérons la classe d'estimateurs $v_L =$

$I' (X' X)^{-1} (\sum_{i=1}^n X_i' A_i r_i' A_i' X_i) (X' X)^{-1} I$, où A_i satis-

fait $A_i' [(I - H_i) U (I - H_i)] A_i = U_i$ pour $i = 1, \dots, n$. Si

$V = K U$ pour un scalaire donné K , alors $E(v_L) = \text{Var}(I'/\beta)$.

Preuve. La valeur attendue de v_L est donnée par

$$E(v_L) =$$

$$= I' (X' X)^{-1} \left(\sum_{i=1}^n X_i' A_i (I - H_i) (K U) A_i' X_i \right) (X' X)^{-1} I$$

$$= I' (X' X)^{-1} \left(\sum_{i=1}^n X_i' A_i (I - H_i) A_i' X_i \right) (X' X)^{-1} I$$

$$= I' (I - H^n) A_i' A_i = I$$

$$A_i' A_i = (I - H^n)^{-1}.$$

Nous posons $U = I$ dans ce qui suit.

Une solution de l'équation (7) existe pour l'UPB i

lorsque $(I - H^n)$ est de plein rang, ce qui est vrai si toutes

les valeurs propres de H^n sont strictement inférieures à 1

(les valeurs propres de H^n sont toujours comprises entre 0

et 1). Une valeur propre de H^n peut être égale à 1 - par

exemple dichotomique dont la valeur est un si, et unique-

ment si, une observation tombe dans la ? UPB.

Si $m_i' < 1$, A_i n'est pas unique. Si A_i satisfait $A_i' A_i =$

$(I - H^n)^{-1}$, alors il en est de même de $O A_i$, pour toute

matrice orthogonale $m_i \times m_i$ dénotée O . Si $V = \sigma^2 I$, le

nous avons $\mathbf{r}_i = (\mathbf{I} - \mathbf{H})^i \mathbf{e}$, où $(\mathbf{I} - \mathbf{H})^i$ contient les m_i lignes de $(\mathbf{I} - \mathbf{H})$ pour la i^{e} UPE. Conséquence,ment,

$$E(v_L) = \left(\frac{n-1}{n} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \left(\sum_{i=1}^l \mathbf{X}'_i (\mathbf{I} - \mathbf{H})^i E(\mathbf{e}) (\mathbf{I} - \mathbf{H})^i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l. \quad (3)$$

$$\sum_{i=1}^l (\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}'_i \mathbf{X} - \mathbf{X}' \mathbf{X}_i + \mathbf{X}' \mathbf{X}) (\mathbf{X}' \mathbf{X})^{-1} l$$

car $E(\mathbf{e}) = \mathbf{I}$ et $(\mathbf{I} - \mathbf{H})^i (\mathbf{I} - \mathbf{H})^i = (\mathbf{I} - \mathbf{H})^{2i}$ pour $\mathbf{H}^{ii} = \mathbf{X}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i$. Soit $\mathbf{D}_i = \mathbf{X}'_i \mathbf{X}_i - \mathbf{X}' \mathbf{X}_i + \mathbf{X}' \mathbf{X}$ = 0. Donc,

$$E(v_L) = \left(\frac{n-1}{n} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \left(\sum_{i=1}^l \mathbf{X}'_i \mathbf{D}_i \mathbf{X}_i \right) l$$

$$\sum_{i=1}^l \left(\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}'_i \mathbf{X} - \mathbf{X}' \mathbf{X}_i + \mathbf{X}' \mathbf{X} \right) l \left[(\mathbf{I} / n) \mathbf{X}' \mathbf{X} + \mathbf{D}_i \right] (\mathbf{X}' \mathbf{X})^{-1} l$$

$$= \left(\frac{n-1}{n} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \left(\mathbf{X}' \mathbf{X} - \mathbf{X}' \mathbf{X} - \mathbf{X}' \mathbf{X} + \mathbf{X}' \mathbf{X} - \sum_{i=1}^l \mathbf{D}_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l$$

$$= l' (\mathbf{X}' \mathbf{X})^{-1} l \left(\frac{n-1}{n} \right) \left(\sum_{i=1}^l \mathbf{D}_i \mathbf{X}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l$$

$$= \text{Var}(l' \beta) - \left(\frac{n-1}{n} \right) \left(\sum_{i=1}^l a'_i (\mathbf{X}' \mathbf{X})^{-1} a_i \right)$$

pour $a_i = \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} l = [\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}' \mathbf{X}_i + \mathbf{X}' \mathbf{X}] (\mathbf{X}' \mathbf{X})^{-1} l$. Comme $(\mathbf{X}' \mathbf{X})^{-1}$ est définie positive, $E(v_L) \leq \text{Var}(l' \beta)$, l'égalité n'ayant lieu que si, et uniquement si, $a_i = 0$, ou de façon équivalente, $\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}' \mathbf{X}_i + \mathbf{X}' \mathbf{X}$ est constante pour tout i .

Les méthodes de répétition ne permettent pas nécessairement d'éviter le problème du biais que produisent les estimateurs de la variance pour la régression. On obtient un estimateur par le jackknife pour les échantillons à plusieurs degrés à partir de l'ensemble de pseudovaleurs $\{\beta^{[i]}\}$, estimations de β d'après des données n^i incluant pas la i^{e} UPE :

$$v_{JK} = [(n-1)/n] \sum_{i=1}^l \left(\beta^{[i]} - \beta \right) \left(\beta^{[i]} - \beta \right) l \quad (5)$$

(Cochran 1977; Rust et Rao 1996). Si $(\mathbf{I}^i - \mathbf{H}^{ii})^{-1}$ existe pour tout i , alors

$$v_{JK} = [(n-1)/n] \sum_{i=1}^l \mathbf{X}'_i (\mathbf{I}^i - \mathbf{H}^{ii})^{-1} \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l^i \quad (6)$$

Par conséquent, une part du biais est proportionnelle à la variance pondérée des moyennes de x au niveau des UPE et une autre, à la variance des sommes des carrés à l'intérieur des UPE.

$$\left\{ \frac{(n-1)M_3^4}{n} \left[\sum_{i=1}^l m_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^l \sum_{j=1}^l m_i m_j (\bar{x}_i - \bar{x})(\bar{x}_j - \bar{x})^2 - m s_2^2 \right] \right\}.$$

La pente est égal à la variance de la variance de la manipulations aléatoires additionnelles montrent que les biais de l'estimateur par linéarisation de la variance de la pente est égal à

$$\frac{m_i}{M s_2^2} \left[(s_2^2 + \bar{x}_2^2 - \bar{x}_i^2) \bar{x}_i - (s_2^2 + \bar{x}_2^2 - \bar{x}_i^2) \bar{x} \right]$$

Pour que v_L et v_{JK} soient non biaisés pour la pente, c'est-à-dire, pour $l' = (0, 1)$, il faut que les expressions $m_i (\bar{x}_i - \bar{x})$ et $m_i (s_2^2 + \bar{x}_2^2 - \bar{x}_i^2)$ soient toutes deux constantes sur tous les i . La première implique que $\bar{x}_i = \bar{x}$, et ensemble, elles impliquent que $m_i s_2^2 = \sum_{j=1}^l (x_{ij} - \bar{x})^2$ est constant. Notons que m_i ne doit pas nécessairement être constant. Cependant, ces deux conditions ne sont pas suffisantes pour garantir que $l' = (0, 1)$, soit sans biais. Des manipulations algébriques additionnelles montrent que le

$$\frac{m_i}{M s_2^2} \left[\bar{x}_i \left(s_2^2 + \bar{x}_2^2 - \bar{x}_i^2 \right) - \bar{x} \left(s_2^2 + \bar{x}_2^2 - \bar{x}_i^2 \right) \right] l$$

où s_2^2 et $\{s_j^2\}$ sont les estimations par maximum de vraisemblance des variances de x , globale et dans les UPE, pour lesquelles les diviseurs sont M et $\{m_i\}$, respectivement. Donc, nous avons

Exemple 1. Considérons la régression linéaire simple. Nous cas de la régression linéaire simple.

L'exemple qui suit montre que les conditions nécessaires pour que les estimateurs par linéarisation et par le jackknife soient non biaisés sont très contraignantes, même dans le cas de la régression linéaire simple.

Théorème 2

Quand $\mathbf{V} = \sigma^2 \mathbf{I}$ et $(\mathbf{I}^i - \mathbf{H}^{ii})^{-1}$ existent pour tout i , alors $E(v_{JK}) \geq \text{Var}(l' \beta)$, l'égalité n'ayant lieu que si, et uniquement si, $l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i$ est constante pour tout i (Preuve en annexe).

2002, page 34). Certains auteurs (Effron et Tibshirani 1993) proposent un autre estimateur par le jackknife où β est remplacé par la moyenne des $\beta^{[i]}$ dans (5). Comme ces deux méthodes produisent des estimations fort semblables dans nos simulations, nous discutons uniquement de la version basée sur (5) dans la suite.

qui découle de la formule mise à jour

Cependant, la méthode de linéarisation a ses limites. Si le nombre d'unités primaires d'échantillonnage est faible, les estimations de l'erreur-type peuvent être fortement biaisées à la baisse, leur coefficient de variation peut être grand et le nombre standard de degrés de liberté pourrait être beaucoup trop grand (Kott 1994; Murray, Haman, Wolfiger, Baker et Dwyer 1998). Par conséquent, l'inférence par linéarisation standard pour les coefficients dont la valeur est fondée en grande partie sur des données provenant d'un petit nombre d'UPB peut produire des intervalles de confiance trop étroits et des tests dont le taux d'erreurs de première espèce est considérablement plus élevé que la valeur nominale. Les méthodes de réutilisation de l'échantillon, comme le jackknife, présentent les mêmes limites.

Dans le présent article, nous caractérisons les facteurs du plan d'échantillonnage (c'est-à-dire, la distribution des variables indépendantes dans les UPB et entre celles-ci) qui produisent un biais important dans les erreurs-types des coefficients de régression linéaire estimées par linéarisation et par le jackknife, et nous démontrons que le problème persiste parfois même si le nombre d'UPB est assez grand. Puis, nous proposons de remplacer l'estimateur par linéarisation standard par un estimateur qui est sans biais en cas d'erreurs indépendantes et idéalement distribués (i.i.d.) et qui a tendance à réduire considérablement le biais autrement. Nous présentons aussi les nombres approximatifs de degrés de liberté à utiliser pour les tests et les intervalles de confiance fondés sur notre estimateur de la variance. Les résultats de simulations montrent que, sur notre test de recchage sont meilleures que celles obtenues par des méthodes plus classiques. Enfin, nous illustrons nos méthodes sur des données provenant d'une expérience nationale conçue pour évaluer le traitement de la dépression.

2. BIAIS DE LA METHODE DE LINEARISATION

Par souci de simplicité, nous limitons l'exposé du corps de l'article à la régression linéaire non pondérée pour des échantillons à deux degrés non stratifiés. Les généralisations aux estimateurs pondérés et aux échantillons stratifiés sont présentées dans McCaffrey, Bell et Bouts (2001) et discutées plus en détail à la section 8.

Soit n , le nombre d'UPB et m_i , le nombre d'unités finales d'échantillonnage provenant de la i^{e} UPB, pour $i = 1, \dots, n$. La taille globale de l'échantillon est $M = \sum_{i=1}^n m_i$. Nous supposons que $y_{ij} = \beta' x_{ij} + \varepsilon_{ij}$, où ε est caractérisée par une moyenne nulle et une matrice de covariances V , et où les y_{ij} , x_{ij} , et ε_{ij} se rapportent tous à la j^{e} observation provenant de la i^{e} UPB. Nous laissons tomber l'hypothèse type des MCO voulant que les erreurs soient i.i.d. et supposons uniquement que les erreurs provenant d'UPB distinctes ne sont pas corrélées. Plus précisément, nous

supposons que V est une matrice diagonale par blocs comptant $m_i \times m_i$ blocs V_i pour $i = 1, \dots, n$. En plus de la notation de ce modèle, dans tout l'article, I représente une matrice d'identité $M \times M$ et I_i une matrice d'identité $m_i \times m_i$. Représentons par β les coefficients estimés du modèle de régression linéaire. Pour simplifier la présentation, nous discutons généralement d'une combinaison linéaire de coefficients de régression, $l' \beta$, pour un vecteur colonne arbitraire l . Dans le cas particulier où un élément de $l = 1$ et les autres sont nuls, $l' \beta$ représente un coefficient estimé unique. Si les erreurs ne sont pas corrélées entre les UPB, la variance de $l' \beta$ est

$$\text{Var}(l' \beta) = l' (X' X)^{-1} \left(\sum_{i=1}^n X_i' V_i X_i \right) (X' X)^{-1} l, \quad (1)$$

où X et X_i sont les matrices de plan d'échantillonnage pour l'échantillon complet et pour l'UPB i , respectivement. L'estimateur par linéarisation standard de la variance de $l' \beta$ est donné par :

$$v_l = l' (X' X)^{-1} \left(c \sum_{i=1}^n X_i' r_i r_i' X_i \right) (X' X)^{-1} l \quad (2)$$

où r_i est le vecteur des résidus pour la i^{e} UPB. En comparant (1) et (2), nous constatons que la linéarisation revient simplement à estimer V_i sous forme du produit d'une constante c par le produit externe des résidus. Habituellement, la valeur de la constante c est fixée à $n/(n-1)$, qui est la valeur utilisée dans SUDAAN et dans les procédures Stata svy (Shah et coll. 1997; StataCorp. 1999). Pour les méthodes EBCG, Zeger et Liang (1986) fixent $c = 1$. Dans des conditions assez générales, m_i converge en probabilité vers la variance de la distribution asymptotique de $\sqrt{n}(l' \beta - l' \beta)$ et le biais relatif de v_l est $O(1/n)$ à mesure que l'augmentation du nombre d'UPB (Fuller 1975; Kott 1994). Pour démontrer la convergence du biais de v_l , Kott (1994) suppose que le nombre d'observations pour chaque UPB est borné et que les éléments de $(X' X)^{-1} X_i'$ sont bornés par b/n pour une constante b . Ces hypothèses assurent effectivement que l'influence de toute UPB sur l'estimation finale diminue à mesure que le nombre d'UPB augmente. La convergence du biais de v_l tient pour les données hétéroscédastiques provenant d'échantillons stratifiés avec poids d'échantillonnage inégaux et une structure de corrélation arbitraire à l'intérieur des UPB. Malheureusement, la convergence ne garantit pas que les propriétés soient bonnes pour un nombre faible à moyen d'UPB.

Théorème 1

Quand $V = \sigma^2 I$ et $c = n/(n-1)$, $E(v_l) \leq \text{Var}(l' \beta)$, l'égalité n'ayant lieu que si, et uniquement si, $l' (X' X)^{-1} X_i' X_j$ est constant pour toute UPB i .

Preuve. Sans perte de généralité, nous supposons que $\sigma^2 = 1$, de sorte que $V = I$. Le vecteur de résidus r peut s'écrire sous la forme $(I - H) e$, où $H = X(X' X)^{-1} X'$ est la matrice chapeau, ou matrice de projection, pour X . Donc,

Réduction du biais des erreurs-types pour la régression linéaire dans le cas d'échantillons à plusieurs degrés

ROBERT M. BELL et DANIEL F. MCCAFFREY¹

RÉSUMÉ

Les méthodes de linéarisation (ou de développement en série de Taylor) sont utilisées très fréquemment pour estimer les erreurs-types des coefficients des modèles de régression linéaire ajustés à des échantillons à plusieurs degrés. Lorsque le nombre d'unités primaires d'échantillonnage (UPB) est grand, la linéarisation peut produire des valeurs précises des erreurs-types dans des conditions assez générales. Par contre, si ce nombre est faible ou que la valeur d'un coefficient dépend en grande partie des données provenant d'un petit nombre d'UPB, les estimateurs par linéarisation peuvent présenter un biais négatif important. Dans le présent article, nous précisons les propriétés de la méthode de plan d'échantillonnage qui biaisent fortement les erreurs-types estimées par linéarisation des coefficients de régression linéaire. Puis, nous proposons une nouvelle méthode, que nous appelons linéarisation à biais réduit (LBR), fondée sur des résidus ajustés afin de mieux approximer la covariance des erreurs réelles. Si les erreurs sont i.i.d., l'estimateur LBR est sans biais pour la variance. En outre, une étude en simulation montre que la LBR peut réduire considérablement le biais, même si les erreurs ne sont pas i.i.d. Nous proposons aussi d'utiliser une approximation de Satterthwaite pour déterminer le nombre de degrés de liberté de la distribution de référence utilisée pour les tests et les intervalles de confiance de combinaisons linéaires de coefficients fondés sur l'estimateur LBR. Nous démontrons que l'estimateur par le jackknife a aussi tendance à être biaisé dans les situations où la linéarisation est biaisée. Cependant, le biais du jackknife est généralement positif. Notre estimateur par linéarisation à biais réduit peut être considéré comme un compromis entre l'estimateur par linéarisation standard et l'estimateur par le jackknife.

MOTS CLÉS : Échantillons complexes; linéarisation; jackknife; approximation de Satterthwaite; degrés de liberté.

1. INTRODUCTION

L'analyse par régression d'échantillons à plusieurs degrés est devenue très courante ces dernières années (par exemple, Ellickson et McGuigan 2000; Shapiro, Morton, McCaffrey, Senierfitt, Fleishman, Perlman, Athey, Keesey, Goldman, Berry et Bozette 1999; Goldstein 1991; Landis, Lepkowski, Eklund et Stehouwer 1982). Même si les modèles hiérarchiques (Bryk et Raudenbush 1992; Gelman, Carlin, Stern et Rubin 1995, chapitre 13) permettent d'analyser à la fois les effets fixes et aléatoires, nombre d'analyses préfèrent la simplicité des modèles de régression classiques quand les effets aléatoires ne présentent pas d'intérêt direct. Les estimateurs classiques de régression produisent des estimations non biaisées des paramètres qui peuvent être efficaces, mais les estimateurs par défaut de l'erreur-type ne tiennent pas compte du plan d'échantillonnage, si bien que les erreurs-types obtenues ne sont pas convergentes (Kish 1965; Skinner 1989a). Diverses méthodes produisent des estimations convergentes des erreurs-types applicables lorsque le nombre d'unités primaires d'échantillonnage (UPB) est suffisamment grand. Elle incluent les méthodes de rééchantillonnage de l'échantillon, comme le jackknife, le bootstrap et les répliques équilibrées répétées, ainsi que les méthodes de linéarisation (ou de

La linéarisation (Skinner 1989b) est une méthode non paramétrique d'estimation des erreurs-types des statistiques fondées sur le plan d'échantillonnage, comme les moyennes et les ratios, ainsi que des coefficients provenant des modèles de régression linéaire et non linéaire. Par non paramétrique, nous entendons que la linéarisation ne repose sur aucune hypothèse quant à la structure de l'erreur à l'intérieur des UPB, comme l'hypothèse d'une corrélation intra-grappe constante. Quand le nombre d'UPB peut être jugé grand, la linéarisation produit des erreurs-types convergentes malgré les caractéristiques multiples des plans d'échantillonnage complexes – stratification, échantillonnage à plusieurs degrés et poids d'échantillonnage – ainsi que des erreurs hétéroscédastiques (Fuller 1975). Étant donné ces propriétés désirables et son intégration de plus en plus fréquente dans des logiciels comme SUDAAN, Stata et SAS Version 8.0 (Shah, Barnwell et Bieler 1997; StataCorp. 1999; SAS Institute, Inc. 1999), la linéarisation est devenue une méthode utilisée couramment pour estimer les erreurs-types et les intervalles de confiance, et pour réaliser des tests statistiques sur des données provenant de plans d'échantillonnage complexes (par exemple, Ellickson et McGuigan 2000; Shapiro et coll. 1999; Rust et Rao 1996). On a aussi proposé la linéarisation pour estimer les erreurs-types produites par les équations d'estimation généralisées (EEG) ajustées à des données à plusieurs degrés (Zegeer et Liang 1986).

¹ Robert M. Bell, Statistics Research Department, AT&T Labs-Research, Room C211, 180 Park Ave., Florham Park, NJ 07932; Daniel F. McCaffrey, Statistics Group, RAND, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516.

STEWART, J. (2000). Alternative Indexes for Comparing Activity Profiles. Article présenté à 2000 International Association for Time-Use Research Conference, Belo Horizonte, Brazil.

TRIPLETT, T. (1995). Data Collection Methods for Estimating Exposure to Pollutants Through Human Activity Pattern Data: A National Micro-behavioral Approach. mimeo, Survey Research Center, University of Maryland.

ensemble de simulations a été étoffé à l'aide de données de la CPS sur les horaires de travail et de données réelles sur l'emploi du temps provenant de l'EPA Time Diary Study de 1992-1994. Les résultats des simulations étoffées confirment ceux des simulations simples et montrent comment le biais peut affecter les estimations du temps consacré à des activités spécifiques. Comme prévu, la stratégie de prise de contact JC introduit un biais d'activité systématique dans les estimations de l'emploi du temps. Le temps consacré que celui consacré aux activités en-dehors de la maison est surestimé. Dans le cas des échantillons produits par les stratégies JDR et JDRS, il n'existe aucun biais d'activité systématique. Les simulations montrent aussi que l'augmentation du nombre de tentatives de prise de contact réduit le biais de non-contact.

Les résultats indiquent clairement que le choix de la stratégie de prise de contact importe et pousse à faire deux recommandations.

Premièrement, les enquêtes sur l'emploi du temps devraient se baser sur un calendrier de prise de contact durant des journées désignées avec possibilité de remise (JDR). Le calendrier JDR crée un biais d'activité plus faible que les autres calendriers de prise de contact pour toutes les hypothèses de profil d'activités testées. Le calendrier JDRS donne des résultats presque aussi bons pour les profils d'activité les plus courants. Toutefois, puisque les taux de contacts et les coûts du travail sur le terrain dépendent du nombre de tentatives de prise de contact, la stratégie JDRS n'est pas plus rentable que la stratégie JDR. Donc, il n'y a aucune raison de préférer la première à la seconde.

Deuxièmement, les enquêtes sur l'emploi du temps doivent inclure des mesures en vue de minimiser le biais de non-contact. Comme ce dernier est en grande partie fonction du nombre de tentatives de prise de contact, un moyen évident de le réduire au minimum consisterait à augmenter le nombre de tentatives de prise de contact. Nous nous étendrons pas davantage sur ce point, car d'autres auteurs ont étudié la question en profondeur. Par exemple, Bauman, Lavradas et Merkle (1993) montrent que l'âge et la situation d'emploi sont liés au nombre de rappels et que des rappels supplémentaires produisent un échantillon plus représentatif et, Bottoman, Masssey et Kalsbeek (1989) proposent une méthode pour déterminer le nombre optimal d'appels. Une autre solution consisterait à essayer d'augmenter la probabilité de rejoindre les répondants potentiels. On pourrait pour cela déterminer quand ils sont susceptibles d'être à la maison et les appeler à ce moment-là, ou leur offrir d'appeler eux-mêmes le jour désigné pour leur interview. Un autre moyen de rendre les répondants potentiels « plus disponibles » consisterait à leur offrir un encouragement monétaire. Une démarche moins coûteuse serait d'ajuster les poids d'échantillonnage. Pothoff et coll. (1993) montrent que, lorsque la variable mesurée est corrélée (entre individus) à la probabilité d'un contact, la

pondération basée sur le nombre de rappels est pratique et efficace. En dernière analyse, la combinaison correcte de ces démarches dépendra des contraintes imposées au gestionnaire de l'enquête.

REMERCIEMENTS

L'auteur remercie John Ellinger, Mike Horrigan, Anne Polivka, Jim Splitzer et Clyde Tucker de leurs commentaires et suggestions. Les opinions exprimées ici sont celles de l'auteur et ne reflètent pas forcément celles du Bureau of Labor Statistics.

BIBLIOGRAPHIE

BAUMAN, S.T., LAVRADAS, P.J. et MERKLE, D.M. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1070-1075.

BOTMAN, S.T., MASSSEY, J.D. et KALSBECK, W.D. (1989). Cost-efficiency and the number of allowable callbacks in the National Health Interview Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 434-439.

HARVEY, A. (1993). Guidelines for time diary data collection. *Social Indicators Research*, 30, 197-228.

HARVEY, A. (1999). Guidelines for time use data collection and analysis. *Time Use Research in the Social Sciences*, (Eds. W.E. Penland, A.S. Harvey, P. Lawton et M.A. McCall). New York: Kluwer Academic/Plenum Publishers, 19-45.

KALTON, G. (1985). Sample design issues in time diary studies. *Time, Goods, and Well-Being*, (Eds. F.T. Juster et F.T. Stafford). Ann Arbor: University of Michigan, Institute of Social Research, 333-351.

KINSLEY, B., et O'DONNELL, T. (1983). Marking time: methodological report of the Canadian time use pilot study-1981. *Explorations in Time Use* (vol. 1), Ottawa: Department of Communications, Employment and Immigration.

LAAKSONEN, S., et PÄÄKKÖNEN, H. (1992). Some methodological aspects on the use of time budget data. *Housework Time in Bulgaria and Finland*, (Eds. L. Kirjavainen, B. Anachkova, S. Laaksonen, I. Niemelä, H. Pääkkönen et Z. Staikevicius), 86-104.

LYBERG, L. (1989). Sampling, nonresponse, and measurement issues in the 1984-85 Swedish Time Budget Survey. *Proceedings of the Fifth Annual Research Conference*, Department of Commerce, Bureau of the Census, 210-238.

POTHOFF, R.F., MANTON, K.G. et WOODBURY, M.A. (1993). Correcting for nonavailability in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, 88, 424, 1197-1207.

STATISTICS CANADA (1999). *Overview of the Time Use of Canadians in 1998*. General Social Survey, Numéro de catalogue 12F0080XIE, Ottawa, Canada.

Tableau 5
Décomposition du biais – Simulations étouffées

	Période de travail sur le terrain de quatre semaines				Période de travail sur le terrain de huit semaines			
	Biais total		Biais de interaction		Biais total		Biais de interaction	
	Biais d'activité	Biais non- contact	Biais d'activité	Biais non- contact	Biais d'activité	Biais non- contact	Biais d'activité	Biais non- contact
Loisirs passifs	JC	-8,62	-7,23	-1,57	0,18	-8,72	-7,29	-1,52
	JD	13,56	0,50	13,16	-0,10	13,51	0,46	13,24
	JDR	2,53	-0,75	3,40	-0,11	-0,35	-0,83	-0,50
	JDRS	0,38	-0,29	0,69	-0,02	-0,31	-0,30	-0,01
Loisirs actifs	JC	4,03	6,27	-1,92	-0,32	4,49	6,80	-1,96
	JD	11,75	-4,40	16,08	0,06	12,30	-3,92	15,97
	JDR	3,31	-1,05	4,15	0,20	0,50	-0,13	0,60
	JDRS	1,08	0,26	0,84	-0,02	0,82	0,83	-0,02
Diversissement/activités sociales	JC	13,11	15,51	-1,89	-0,51	13,06	15,53	-1,92
	JD	15,78	1,30	15,82	-1,34	15,80	1,32	15,69
	JDR	5,64	1,72	4,08	-0,17	2,47	1,91	0,59
	JDRS	1,37	0,58	0,82	-0,04	0,40	0,42	-0,02
Activités organisationnelles	JC	15,24	17,36	-1,70	-0,42	14,89	17,06	-1,76
	JD	12,26	2,05	14,28	-1,08	14,88	1,72	14,39
	JDR	12,37	8,30	3,69	0,39	7,14	6,53	0,54
	JDRS	5,99	5,24	0,74	0,01	4,76	4,77	-0,02
Éducation/formation	JC	19,17	22,84	-2,49	-1,18	19,73	23,53	-2,56
	JD	22,02	1,94	20,90	-0,82	22,74	2,54	20,90
	JDR	13,39	9,04	5,40	0,96	10,29	9,36	0,78
	JDRS	8,00	6,78	1,09	0,13	7,32	7,35	-0,02
Soins personnels	JC	-0,79	-0,51	-0,28	0,00	-0,82	-0,53	-0,29
	JD	2,20	-0,13	2,39	-0,06	2,20	-0,13	2,39
	JDR	0,34	-0,26	0,62	-0,02	-0,17	-0,26	0,09
	JDRS	-0,15	-0,27	0,12	0,00	-0,29	-0,29	0,00
Achats de biens/services	JC	4,67	7,55	-2,39	-0,49	4,48	7,44	-2,45
	JD	22,36	2,34	20,06	-0,04	22,23	2,23	20,00
	JDR	4,25	-1,02	5,18	0,10	-0,42	-1,18	0,75
	JDRS	-1,49	-2,54	1,04	0,01	-2,58	-2,55	-0,02
Garde active d'enfants	JC	-9,09	-7,81	-1,40	0,11	-9,14	-7,82	-1,43
	JD	14,21	0,45	11,72	2,04	14,30	0,33	11,66
	JDR	0,77	-2,32	3,03	0,07	-2,23	-2,69	0,44
	JDRS	-0,09	-0,69	0,61	0,00	-0,89	-0,87	-0,01
Taux ménagers	JC	-11,49	-9,42	-2,26	0,18	-11,64	-9,51	-2,32
	JD	20,77	1,43	18,93	0,41	20,63	1,31	18,95
	JDR	4,53	-0,43	4,88	0,08	-0,17	-0,55	0,71
	JDRS	2,52	1,50	0,99	0,03	1,34	1,36	-0,02
Travail rémunéré	JC	6,74	2,95	3,69	0,11	6,86	2,96	3,79
	JD	-31,43	-0,77	-30,90	0,24	-31,44	-0,78	-30,90
	JDR	-7,74	0,25	-7,98	-0,02	-0,86	0,30	-1,16
	JDRS	-1,87	-0,27	-1,61	0,00	-0,22	-0,26	0,03

simulations montrent aussi que les estimations obtenues en utilisant un calendrier de prise de contact à journées convenables (JC) sont sensibles à la variance de la probabilité individuelle (au niveau de la personne) d'un contact. Les estimations du temps consacré aux activités qui sont positivement corrélées à la probabilité d'un contact (par exemple, activité ayant lieu à la maison) diminuent à mesure qu'augmente la variance. En revanche, les estimations produites par d'autres calendriers de prise de contact ne sont pas sensibles à la variance individuelle de la probabilité d'un contact. Étant donné les résultats des simulations simples, il est évident que le biais global calculé pour les diverses stratégies de prise de contact dépend de la fréquence relative de chaque profil dans la population. Comme il n'existe pas de données directes sur ces profils, le premier

Nous pouvons tirer plusieurs enseignements de ces décompositions (présentées au tableau 5). Premièrement, dans le cas du calendrier JC, le biais global est dû entièrement au biais d'activité. Le grand nombre de tentatives de prise de contact garantissant virtuellement l'obtention d'un échantillon représentatif, si bien qu'au augmentant la période de travail sur le terrain de quatre à huit semaines n'apporte que peu de changements. Par contre, le biais de non-contact représente la totalité du biais dans le cas du calendrier JD. Dans le cas des calendriers JDR et JDRS, il n'existe virtuellement aucun biais d'activité et le biais de non-contact diminue spectaculairement lorsque la période de travail sur le terrain passe de quatre à huit semaines. Il n'est pas étonnant de constater que pour le calendrier JDR, le biais de non-contact associé à une période de travail sur le terrain de huit semaines est à peu près le même que le biais de non-contact obtenu pour le calendrier JDRS avec travail sur le terrain de quatre semaines. Dans ces simulations, l'échantillon devient entièrement représentatif lorsque la période de travail sur le terrain est suffisamment longue pour permettre 16 tentatives de prise de contact. Enfin, la faible importance des termes d'interaction témoigne du fait que les biais d'activité et de non-contact associés à chaque stratégie de prise de contact sont corrélés négativement.

4. SOMMAIRE ET RECOMMANDATIONS

Les enquêtes téléphoniques sur l'emploi du temps ont des caractéristiques uniques qui compliquent la collecte des données. Contrairement à la plupart des autres enquêtes, celles sur l'emploi du temps ne peuvent s'appuyer sur des réponses par procuration, de sorte qu'il est plus probable que la probabilité de rejoindre un répondant potentiel soit corrélée aux activités de ce dernier. En outre, comme, dans le cas des enquêtes téléphoniques sur l'emploi du temps, on demande aux répondants de déclarer leur emploi du temps de la journée précédente, il se peut que la probabilité d'interviewer le répondant au sujet d'une journée de référence particulière soit corrélée aux activités ayant eu lieu durant cette journée de référence. Nous montrons d'abord comment ces caractéristiques peuvent donner lieu à un biais de non-contact et à un biais d'activité. Puis, à l'aide de deux ensembles de simulations informatiques, nous montrons que l'importance de ces biais dépend de la stratégie adoptée pour prendre contact avec les répondants potentiels.

Le premier ensemble de simulations montre que l'importance du biais associé à un calendrier particulier de prise de contact dépend du profil des journées à prise de contact (PCD). Le calendrier de prise de contact à journées désignées avec remise (JDR) donne de meilleurs résultats que les autres calendriers pour tous les profils d'activité étudiés. Ces

passé de quatre à huit semaines. L'utilisation d'un calendrier de prise de contact JDRS durant une période de travail sur le terrain de huit semaines (16 tentatives de prise de contact) produit un taux de contacts de 80 % et un échantillon représentatif. Évidemment, l'échantillon obtenu pour le calendrier JDR avec une période de travail sur le terrain de huit semaines est virtuellement identique à celui obtenu avec un calendrier JDRS et une période de travail sur le terrain de quatre semaines.

Biais d'activité c. biais de non-contact

Pour obtenir une image plus précise de la contribution de chaque type de biais au biais global, nous avons décomposé ce dernier en une fraction due au biais d'activité, une fraction due au biais de non-contact et une fraction due à l'interaction entre les deux biais. Pour l'activité a et le groupe g (travailleurs ou non-travailleurs), le biais global est donné par :

$$F_g^a(X_{ag}^* - X_{ag}^*) + X_{ag}^*(F_g^a - F_g^*) + (F_g^a - F_g^*)(X_{ag}^* - X_{ag}^*) + \text{Interaction}$$

Tableau 4

Sommaire des taux de contact – Simulations étoffées

Période sur le terrain		Valeur	
4 semaines	Taux de contacts	JC	JDR
		89,68	40,35
	Pourcentage de non-travailleurs	40,08	60,07
	Pourcentage de travailleurs	59,92	39,93
8 semaines	Taux de contacts	89,79	40,35
		88,17	78,87
	Pourcentage de non-travailleurs	42,19	42,88
	Pourcentage de travailleurs	57,81	57,12

où F_g^a est la fraction de l'échantillon dans le groupe g, et X_{ag}^* est le temps consacré à l'activité a par le groupe g, et les astérisques indiquent les valeurs réelles. On obtient le biais total pour l'activité a par sommation de cette expression sur les travailleurs et les non-travailleurs, c'est-à-dire :

$$\sum_{g=W,N} (F_g^a X_{ag}^* - F_g^a X_{ag}^*) = \sum_{g=W,N} F_g^a (X_{ag}^* - X_{ag}^*) + \sum_{g=W,N} X_{ag}^* (F_g^a - F_g^*) + \sum_{g=W,N} (F_g^a - F_g^*)(X_{ag}^* - X_{ag}^*)$$

ont lieu en-dehors de la maison (loisirs actifs, divertissements/activités sociales, activités organisationnelles, éducation/formation, achat de biens/services et travail non rémunéré) et négatif pour les activités qui ont lieu à la maison (loisirs passifs, soins personnels, garde active d'enfants et travaux ménagers). Ce profil confirme le résultat des études mentionnées dans l'introduction, à savoir que le temps consacré à des activités en-dehors de la maison devient plus représentatif à mesure que le nombre de tentatives de prise de contact augmente (voir le tableau 4). C'est pour la stratégie JD que le taux de contacts est le plus faible (40 %) et que l'échantillon est le moins représentatif. Dans le cas des calendriers de prise de contact JDR et JDRS, le taux de contacts augmente et l'échantillon devient plus représentatif lorsque la période de travail sur le terrain

est plus important dans le cas d'une stratégie de prise de contact durant des journées convenables pour le répondant que dans celui d'une stratégie à prise de contact lors de journées désignées. Mais avant tout, il est maintenant évident que ce résultat est dû à un biais lié à la stratégie de

Tableau 3b

Biais estimatif – Simulations étouffées (période de travail sur le terrain de huit semaines)

Activité/Situation d'emploi	JC	JD	JDR	JDRS	Temps consacré à l'activité (réel)
Loisirs passifs	-8,63*	-0,09	-1,62	-1,21	315,38
Non-travailleurs	-5,24*	1,28	0,39	1,10	151,72
Global	-8,72*	-13,51*	-0,35	-0,31	220,79
Loisirs actifs	10,62*	-2,03	1,76	0,06	65,46
Non-travailleurs	0,00	-7,29	-3,50	2,21	26,87
Global	4,49*	12,30*	0,50	0,82	43,16
Diversissements/activités sociales	19,77*	-1,72	-0,15	-0,91	67,10
Non-travailleurs	8,09*	6,64	5,52	2,76	28,00
Global	13,06*	15,80*	2,47	0,40	44,50
Activités organisationnelles	18,92*	-1,53	8,59	3,25	19,36
Non-travailleurs	14,03*	7,00	3,18	7,25	8,72
Global	14,89*	14,88*	7,14*	4,76	13,21
Education/formation	33,56*	0,18	12,91*	9,55*	43,34
Non-travailleurs	-0,72	8,24	0,77	2,01	13,09
Travailleurs	19,73*	22,74*	10,29*	7,32*	25,86
Soins personnels	-0,50	-0,29	-0,48	-0,44	663,03
Non-travailleurs	-0,55*	0,00	-0,08	-0,16	580,81
Global	-0,82*	2,20*	-0,17	-0,29	615,51
Achats de biens/services	12,64*	1,36	-0,09	-1,28	72,97
Non-travailleurs	-4,41	4,23	-3,66	-5,45*	23,36
Global	4,48*	22,23*	-0,42	-2,58	44,30
Garde active d'enfants	-7,67*	5,36	-1,04	-0,31	24,07
Non-travailleurs	-8,02*	-6,18	-4,98	-1,65	12,66
Global	-9,14*	14,30*	-2,23	-0,89	17,48
Travaux ménagers	-9,02*	1,55	0,20	2,10	169,30
Non-travailleurs	-10,55*	0,80	-2,15	-0,20	57,95
Global	-11,64*	20,63*	0,17	1,34	104,94
Travail rémunéré	—	—	—	—	—
Non-travailleurs	2,96*	-0,78	0,30	-0,26	536,82
Global	6,86*	-31,44*	-0,86	-0,22	310,25

Tableau 3a
Biais estimatif – Simulations étouffées (période de travail sur le terrain de quatre semaines)

Activité/Situation d'emploi	JD	JDR	JDRS	Temps consacré à l'activité (réel)
Loisirs passifs				
Non-travailleurs	-8,44 *	-1,54	-1,03	314,72
Travailleurs	-5,40 *	0,43	0,82	152,04
Global	-8,62 *	2,53 *	0,38	220,70
Loisirs actifs				
Non-travailleurs	9,80 *	-2,75	-0,99	65,94
Travailleurs	-0,07	-7,34	-4,69	26,89
Global	4,03 *	11,75 *	3,31	43,37
Diversissement/activités sociales				
Non-travailleurs	19,41 *	-2,01	-0,25	67,30
Travailleurs	8,63 *	7,14	5,21	27,87
Global	13,11 *	15,78 *	5,64 *	44,51
Activités organisationnelles				
Non-travailleurs	19,58 *	-0,98	9,00	19,25
Travailleurs	13,77 *	6,95	7,17	8,72
Global	15,24 *	15,26 *	12,37 *	13,16
Education/formation				
Non-travailleurs	32,77 *	-0,42	12,54 *	43,60
Travailleurs	-1,17	7,63	0,57	13,16
Global	19,17 *	22,02 *	15,39 *	26,01
Soins personnels				
Non-travailleurs	-0,50	-0,29	-0,49	663,04
Travailleurs	-0,52 *	0,01	-0,06	580,71
Global	-0,79 *	2,20 *	0,34	615,46
Achats de biens/services				
Non-travailleurs	12,62 *	1,35	0,11	72,98
Travailleurs	-4,05	4,62	-3,62	23,28
Global	4,67 *	22,36 *	4,25 *	44,25
Garde active d'enfants				
Non-travailleurs	-7,89 *	5,11	-1,06	24,13
Travailleurs	-7,69 *	-6,05	-4,09	12,64
Global	-9,09 *	14,21 *	0,77	17,49
Travaux ménagers				
Non-travailleurs	-8,88 *	1,71	0,33	169,04
Travailleurs	-10,55 *	0,85	-2,03	57,92
Global	-11,49 *	20,77 *	4,53 *	104,82
Travail rémunéré				
Non-travailleurs	—	—	—	—
Travailleurs	2,95 *	-0,77	0,25	536,77
Global	6,74 *	31,44 *	-7,74 *	310,22

Nota : L'astérisque indique que le biais dans l'estimation du temps consacré à l'activité diffère significativement de zéro au niveau de signification de 5 %.

nous le montrerons plus bas, est due principalement au biais

de non-contact.

La comparaison des trois autres stratégies de prise de contact permet de dégager deux tendances. En premier lieu, le biais d'activité est nettement plus faible (et généralement statistiquement non significatif) si l'on utilise la stratégie JDR ou la stratégie JDRS plutôt que la stratégie JC. En deuxième lieu, le biais qui entache les estimations JC suit le profil prévu. Il a tendance à être positif pour les activités qui

La stratégie JD ne donne lieu à virtuellement aucun biais d'activité. Pour quelques activités – loisirs actifs, diversissement/activités sociales, activités organisationnelles, éducation/formation et garde active d'enfants pour les travailleurs, et garde active d'enfants pour les non-travailleurs – le biais d'activité est assez important, mais aucune des estimations n'est statistiquement significative. Dans le cas de la stratégie JD, le biais global est assez important pour la plupart des activités, situation qui, comme

Pour calculer les probabilités d'un contact, il a fallu émettre une troisième hypothèse. À l'instar de Pothoff, Manton et Woodbury (1993), nous avons supposé que la probabilité d'un contact était égale au nombre de minutes consacrées à une activité effectuée à la maison (sauf dormir) divisé par le temps consacré à toutes les autres activités que le sommeil. Ce processus de calcul de la probabilité d'un contact présente deux propriétés importantes : (1) la probabilité d'un contact pour une journée donnée est liée aux activités réalisées cette journée-là et (2) l'un des groupes de répondants potentiels (travailleurs) a une probabilité plus faible de contact productif (0,36 c. 0,72). Les tableaux 3a et 3b résument les estimations du biais

travailleurs et les non-travailleurs, et une estimation du biais global. Comme ce dernier inclut le biais de non-contrat, il est possible qu'il soit plus grand (ou plus petit) que le biais d'activité pour l'un ou l'autre groupe. Nous avons calculé le biais de la même façon que pour la simulation précédente. Comme pour l'ensemble précédent de simulations, un astérisque indique que le biais est significativement nul au niveau de signification de 5 %. La cinquième colonne montre le temps réellement consacré à chaque

designés (JD) pour laquelle la période de travail sur le terrain n'a pas d'importance, la différence principale est que le biais global est plus petit lorsque la période de travail sur le terrain est de huit semaines. Ce biais global plus faible est dû principalement au nombre accru de tentatives de prise de contact qui fait augmenter de façon disproportionnée la probabilité que les travailleurs soient rejoints et rend l'échantillon plus représentatif (voir le tableau 4). En revanche, les estimations du biais d'activité associée aux diverses stratégies de prises de contact ne sont pas sensibles à la longueur de la période durant laquelle sont faites les tentatives de contact. Le reste de la discussion se concentrera sur les résultats du tableau 3b.

Distribution des horaires de travail

Profil d'activité	Pourcentage	Pourcentage cumulé
L	-	39,40
Ma	-	87,51
Me	-	90,14
J	-	91,77
V	-	92,58
	-	92,84
	-	93,21
	-	93,89
	-	94,38
	-	94,63
	-	95,14
	-	95,39
	-	96,12
	-	96,48
	-	97,18
	-	97,82
	-	100,00

Pour produire des renseignements pour les activités individuelles, nous nous sommes servis de données provenant de l'EPA Time Diary Study de 1992-1994 réalisé par l'Université du Maryland. Cet ensemble de données contient des journaux sur l'emploi du temps pour un ensemble de 7 408 adultes (voir Triplett 1995). Comme n'existe qu'une seule observation par personne. Nous avons utilisé la méthode d'échantillonnage répétée qui suit pour produire des données correspondant à huit semaines pour un échantillon de 18 974 "individus". Nous avons réparti les journées du journal entre les journées de travail et les journées de congé. Nous avons considéré qu'une journée du journal correspondait à une journée de travail si l'individu avait fait un travail rémunéré durant la journée en question. Les journées de travail ont été attribuées aux travailleurs et les journées de congé ont été attribuées aux non-travailleurs. Les lundi ont été tirés des observations faites le lundi, les mardi, des observations faites le mardi, etc.

sauf le calendrier à journées convenables (JC). Comme prévu, le calendrier JC donne lieu à une surestimation de l'indice d'activité moyenne. Mais, par-dessus tout, lorsqu'on utilise le calendrier JC, l'indice estime d'activité lorsque les activités ne sont pas corrélées d'une journée à l'autre – est corréliée positivement à la variance de ε . Puisque la variance augmente de 0,003 ($\varepsilon = 0,1$) à 0,083 ($\varepsilon = 0,5$), le biais augmente pour passer de moins de 1 % à 15 %. On peut voir ce que signifie intuitivement ce résultat en notant qu'une réalisation négative importante de ε pour une journée particulière rend les répondants durant une fraction assez importante des journées PCF et, donc, les journées du journal comptent un nombre disproportionné de journées PCD. Naturellement, si l'on modifie le calendrier JDRS de sorte que les tentatives de contact aient lieu les deux mêmes journées chaque semaine, le biais est virtuellement nul.

Ces simulations montrent clairement que le biais d'activité associé à chaque stratégie de prise de contact dépend du profil journalier des activités, des probabilités d'un contact durant une journée PCD ou une journée PCF, et de la variance de ces probabilités. Cependant, il est également évident que le calendrier JDR donne de meilleurs résultats que les autres, quel que soit le profil d'activité supposé. Si l'on considère chaque profil comme une catégorie différente de répondants, alors le biais global (qui inclut à la fois le biais d'activité et le biais de non-contact) dépend de la fréquence relative de chaque catégorie dans la population. Des renseignements sur l'incidence de chaque catégorie permettrait de mesurer le biais global et, pour chaque stratégie, de décomposer ce dernier en une fraction due au biais d'activité et une fraction due au biais de non-contact. Nous examinons cette question à la section suivante.

Profil 2 à 7 – Probabilités de base groupées

Les résultats sont mixtes lorsque les journées PCD sont groupées soit au début soit à la fin de la semaine. Dans les simulations où la différence $P_P - P_D$ est assez faible (0,2), tous les calendriers de prise de contact donnent d'assez bons résultats. La valeur absolue du biais est inférieure à 3 % dans tous les cas. Cependant, si $P_P - P_D$ est assez grande (0,5), les écarts entre les biais associés, respectivement, à chaque calendrier de prise de contact sont significatifs. Le calendrier JDR est celui qui donne les meilleurs résultats dans l'ensemble. Le biais n'excède 5 % (en valeur absolue) que pour le profil 7 (FFFD), pour lequel il est égal à -5,5 %. Par contre, si l'on utilise les calendriers JD et JDRS, le biais est de l'ordre de 10 à 14 % pour les profils 2 (DDFF) et 3 (DDFF), et de l'ordre de 16 à 20 % pour les profils 6 (FFDD) et 7 (FFFD). La différence entre les calendriers JD et JDRS, d'une part, et le calendrier JDR, d'autre part, pour ces profils est significative, statistiquement et en pratique. Pour les profils 4 (DDDF) et 5 (FFDD), le calendrier JDR donne des résultats un peu moins bons que les calendriers JD et JDRS, mais le biais est si faible (inférieur à 1,5 %) que la différence n'a aucune signification pratique. Le calendrier JC donne d'un peu meilleurs résultats que les calendriers JD et JDRS. Le biais est de l'ordre de 1 à 18 %. Comme pour le profil 1 susmentionné, l'indice estime d'activité moyenne augmente avec la variance de ε pour le calendrier JC, mais non les autres. Et, comme le montre le tableau 1, pour les profils où le biais est négatif (profils 6 et 7), une augmentation de la variance de ε réduit le biais.

Profil 8 – Probabilités de prise de contact alternées

Tous les calendriers de prise de contact produisent des estimations biaisées, car les journées PCF sont des échantillons biaisés. Comme plus haut, tous les calendriers donnent d'assez bons résultats si la différence $P_P - P_D$ est assez faible. Le biais est de l'ordre de 5 à 8 % pour tous les calendriers, sauf le calendrier JDR, pour lequel il est d'environ 1 %. Cependant, si $P_P - P_D$ est importante, tous

3. SIMULATIONS ÉTOFFÉES

Si l'on est prêt à émettre certaines hypothèses supplémentaires, il est possible d'étroffer les simulations à l'aide de données provenant d'autres sources. La première hypothèse est que les horaires de travail des individus constituent une approximation raisonnable des profils d'activité des journées PCD et PCF, de sorte que les journées de travail correspondent aux journées PCD et les journées de congé, aux journées PCF. La deuxième hypothèse est qu'il est possible de reproduire la semaine d'un individu donné en prenant une journée de la semaine de cinq individus distincts.

Nous avons utilisé les données du Work Schedule Supplement de mai 1997 à la Current Population Survey (CPS) pour obtenir des renseignements sur les horaires de travail individuels. Notons que, comme il faut connaître la prévalence de chaque type d'horaire pour la population complète, nous avons également inclus les personnes qui ne travaillent pas. Le tableau 2 donne les profils des journées de travail (W) et de congé (N) basés sur la CPS de mai 1997. Deux de ces profils englobent environ 88 %

8. Pour la moitié de l'échantillon, les lundi, mercredi et vendredi sont des journées PCD et les mardi et jeudi, des journées PCF (PFDPD). Pour l'autre moitié de l'échantillon, la situation est inversée (FDPDF).

Dans le cas du profil 1, la probabilité de base de joindre le répondant est la même dans tous les cas, si bien que la variation des probabilités est due entièrement aux termes aléatoires. Dans le cas des profils 2 à 7, les journées PCD sont regroupées soit au début soit à la fin de la semaine. Dans le cas du profil 8, les probabilités de base alternent entre celles des journées PCD et PCF.

Pour étudier le biais d'activité, nous avons exécuté des simulations distinctes pour chacun des huit profils décrits plus haut. Donc, pour une simulation particulière, tous les individus affichent le même profil de probabilités de base. Le tableau 1 donne les résultats provenant d'un sous-ensemble représentatif des 153 simulations exécutées. Les

Tableau 1 Biais d'activité associé à chaque stratégie de prise de contact sous diverses hypothèses concernant la corrélation des activités

Profil d'activités	Probabilité moyenne d'un contact		Biais estimé (exprimé en pourcentage de l'indice d'activité réelle)	Biais estimé (exprimé en pourcentage de l'indice d'activité réelle)			
	Journées à contact difficile	Journées à contact facile	ε	Indice d'activité réelle	JC	JD	JDR

Probabilités de base identiques	0,50	0,50	0,10	0,50	0,7*	0,7*	0,1
DDFFF	0,75	0,25	0,05	0,500	0,7	-10,7*	-4,7*
	0,60	0,40	0,05	0,500	5,2*	-10,9*	-4,8*
	0,75	0,25	0,05	0,500	-0,1	-0,7*	-2,8*
	0,60	0,40	0,20	0,500	-2,2*	-0,7*	-2,5*
	0,75	0,25	0,05	0,550	1,9*	-2,4*	-0,5
	0,60	0,40	0,20	0,550	-0,4*	-1,8*	-0,6*
	0,75	0,25	0,05	0,625	0,8	-10,3*	-4,1*
	0,60	0,40	0,25	0,625	-2,7*	-9,7*	-4,0*
	0,75	0,25	0,05	0,625	-2,5*	-2,6*	-0,7*
	0,60	0,40	0,20	0,500	-0,1	-0,7*	-2,5*
	0,75	0,25	0,05	0,500	5,2*	-10,9*	-4,8*
	0,60	0,40	0,25	0,500	-0,1	-0,7*	-2,8*
	0,75	0,25	0,05	0,750	0,1	-0,1	0,0
	0,60	0,40	0,25	0,750	2,3*	-0,5	0,2
	0,75	0,25	0,05	0,600	1,9*	-0,3	0,2
	0,60	0,40	0,20	0,600	1,7*	1,0	1,4*
	0,75	0,25	0,25	0,625	4,2*	-0,3	1,2*
	0,60	0,40	0,20	0,550	1,1*	0,3	0,3
	0,75	0,25	0,05	0,500	-18,2*	-17,1*	-4,3*
	0,60	0,40	0,25	0,500	-15,9*	-17,9*	-4,5*
	0,75	0,25	0,05	0,500	-2,0*	-2,2*	-0,4
	0,60	0,40	0,20	0,500	-0,4	-2,4*	-2,6*
	0,75	0,25	0,05	0,375	-16,6*	-17,6*	-5,5*
	0,60	0,40	0,25	0,450	-11,4*	-17,6*	-5,6*
	0,75	0,25	0,05	0,500	31,5*	26,4*	9,6*
	0,60	0,40	0,25	0,500	34,7*	25,5*	9,7*
	0,75	0,25	0,05	0,500	7,8*	4,5*	1,3*
	0,60	0,40	0,20	0,500	1,2*	4,3*	1,2*
	0,75	0,25	0,05	0,500	28,5*	26,4*	9,6*
	0,60	0,40	0,25	0,500	29,4*	25,5*	9,7*
	0,75	0,25	0,05	0,500	5,1*	4,5*	1,3*
	0,60	0,40	0,20	0,500	5,1*	4,3*	1,2*

Nota : Les astérisques indiquent que l'indice estimé d'activité moyenne est statistiquement différent de la valeur réelle au niveau de signification de 5 %.

contact de la première semaine soient les mardi/jeu-di ou mercredi/ven-dredi dépend de la journée de départ, qui est attribuée au hasard.

Comme l'illustre l'exemple 1, il est facile de montrer que l'activité dans les estimations de l'emploi du temps si la probabilité de base d'un contact est la même pour chaque journée (0,5), à part le bruit aléatoire (+0,5 avec une probabilité de 1/2 ou -0,5 avec une probabilité de 1/2). Même si Stewart (2000) montre que les journées comprises entre le lundi et le jeudi sont fort semblables en moyenne, il est probable que, pour certaines personnes, les probabilités d'un contact varient systématiquement selon le jour de la semaine. Par exemple, il pourrait être difficile de rejoindre certaines personnes les lundi, mercredi et ven-dredi de chaque semaine. Cette variation systématique rend consi-dérablement plus difficile le dépistage d'un biais dans les estimations d'échantillon, ainsi que la détermination de la direction et de la grandeur de ce biais. On pourrait modé-liser des stratégies de prise de contact et résoudre analyti-quement pour le biais sous diverses hypothèses au sujet du profil des probabilités d'un contact. Il s'agit là toutefois d'un processus fastidieux, car chaque hypothèse concernant le profil des probabilités d'un contact selon la journée nécessiterait une solution distincte. Par contre, les simula-tions informatiques constituent un moyen idéal d'évaluer le biais associé à diverses stratégies de prise de contact sous diverses hypothèses quant au profil des probabilités d'un contact. Le programme informatique est plus simple et produit des résultats plus intuitifs que la solution analytique. En outre, il est très facile de le modifier en vue d'utiliser des profils différents. À la section 3, nous ajoutons une touche de réalisme aux simulations en intégrant des données réelles sur l'emploi du temps – ce qui serait impossible si on adoptait la démarche analytique.

Simulations

La stratégie de simulation est très simple. En premier lieu, nous avons créé, pour chacun des 10 000 répondants potentiels, des « données » pour une période de quatre semaines. Afin de nous concentrer sur les stratégies de prise de contact, nous ne nous sommes pas préoccupés des méthodes d'échantillonnage et avons supposé que l'échan-tillon de répondants potentiels était représentatif de la population. Les simulations sont conçues pour permettre de comparer les calendriers de prise de contact susmentionnés, de sorte que nous supposons que la « semaine » compte cinq jours. Nous avons limité les journées admissibles pour produire le journal de l'emploi du temps du lundi au jeudi, car, comme nous l'avons mentionné plus haut, ces journées sont celles qui se ressemblent le plus. L'étape suivante consistait à simuler les tentatives de prise de contact avec ces répondants conformément aux quatre calendriers de prise de contact décrits plus haut. Enfin, nous avons comparé les estimations ainsi obtenues aux valeurs d'échantillon.

Pour simplifier les simulations, nous avons produit un résumé analytique d'activités spécifiques, comme dans les exemples qui précèdent, et caractérisé chaque journée à l'aide d'un indice d'activité, I_j , ($j = D, F$) qui varie de 0 à 1. Cet indice d'activité est donné par $I_j = 1 - P_j$ où P_j représente la probabilité de rejoindre et d'interviewer le répondant. Pour simuler la variation des activités d'une journée à l'autre, nous avons posé que la probabilité d'un contact lors d'une journée particulière est donnée par :

$$P_j = P_j + \varepsilon_j$$

où P_j est la probabilité moyenne d'un contact durant une journée PCD ($j = D$) ou PCF ($j = F$), et $\varepsilon \sim U(-\varepsilon, \varepsilon)$. Nous supposons que $P_D > P_F$, autrement dit, en moyenne, les répondants sont moins susceptibles d'être rejoints durant les journées PCD que durant les journées PCF. Pour nous assurer que les probabilités d'un contact soient comprises dans l'intervalle $[0, 1]$, nous établissons ε de sorte que $\varepsilon < \min(P_D, 1 - P_F)$.

Nous pouvons formuler de nombreuses hypothèses concernant le profil des activités selon le jour de la semaine. Le cas le plus simple est celui où toutes les journées sont identiques, sauf en ce qui concerne le bruit aléatoire. Toutefois, comme nous l'avons mentionné plus haut, il se peut qu'il soit systématiquement plus difficile de joindre les répondants potentiels certains jours que d'autres. Pour couvrir une grande gamme de profils d'activités, nous avons effectué les simulations sous les huit hypothèses qui suivent en ce qui concerne la répartition des journées PCD et PCF durant chacune des quatre semaines.

1. Les valeurs réelles de l'indice d'activité sont distribuées comme $U(0, 1)$, de sorte que la valeur moyenne est 0,5.
2. Les deux premières journées de chaque semaine sont PCF (DDFF).
3. Les trois premières journées de chaque semaine sont PCF (DDFF).
4. Les quatre premières journées de chaque semaine sont des journées PCD et la dernière est une journée PCF (DDDDF).
5. La première journée de chaque semaine est une journée PCD et les quatre dernières sont des journées PCD (FDDDD).
6. Les deux premières journées de chaque semaine sont des journées PCF et les trois dernières, des journées PCD (FFDD).
7. Les trois premières journées de chaque semaine sont des journées PCF et les deux dernières, des journées PCD (DDFF).

prise de contact sont faites durant des journées PCD et l'autre moitié, durant des journées PCF, l'indice d'activité moyenne pour l'échantillon final est égal à $0,75 (= 0,5 \times 0,5 + 0,5 \times 1)$.

Exemple 2 – Biais de non-contact : Supposons maintenant que les répondants potentiels diffèrent l'un de l'autre en ce qui a trait aux probabilités d'un contact et que, pour chaque personne, cette probabilité ne varie pas d'un jour à l'autre. Supposons aussi que la moitié des répondants potentiels sont rejoints durant des journées PCD, avec la probabilité d'un contact $P_D = 0,25$, et que l'autre moitié sont rejoints durant des journées PCF avec probabilité d'un contact $P_F = 0,75$. Si nous essayons de contacter chaque répondant potentiel quatre fois, étant donné ces probabilités, pratiquement tous les répondants potentiels PCF (99,6 %) seront rejoints. En revanche, 68,4 % seulement des répondants potentiels PCD le seront. Le taux global de contacts sera de 84 % ($99,6 \% \times 0,50 + 68,4 \% \times 0,50$), mais l'échantillon final ne sera pas représentatif, puisque 59,3 % seront des répondants PCF et 40,7 % seulement des répondants PCD. Par conséquent, les estimations basées sur cet échantillon auront tendance à sous-estimer le temps consacré aux activités entreprises par les personnes PCD et à surestimer celui consacré aux activités entreprises par les personnes PCF.

Les biais décrits plus haut ne se produisent pas uniquement dans le cas des enquêtes sur l'emploi du temps. Dans la plupart des enquêtes, des mesures sont prises pour réduire au minimum le biais de non-contact, mais moins d'attention est accordée au biais d'activité. Par exemple, outre l'objectif principal consistant à recueillir des données biographiques sur l'emploi, les enquêtes nationales longitudinales incluent quelques questions sur les activités de la population active (emploi et heures de travail) durant la semaine qui a précédé l'interview. Comme le moment de l'interview est généralement fixé à la meilleure convenance du répondant, les activités de ce dernier durant la semaine de référence seront corrélées à la probabilité de l'interviewer au sujet de cette semaine de référence. Le raisonnement intuitif qui sous-tend cette corrélation est exactement le même que pour l'exemple 1. Cette corrélation introduit un biais dans les estimations du nombre d'heures de travail, mais la direction de ce biais est indéterminée. Le nombre d'heures travaillées par semaine a tendance à être surestimé chez les personnes pour lesquelles on a dû repousser l'interview à cause d'un horaire de travail chargé et à être sous-estimé chez celles qui étaient en congé. Le biais d'activité pose aussi un problème dans le cas des enquêtes sur les voyages. Le temps passé en-dehors de la maison a tendance à être surestimé si l'on demande aux répondants des renseignements au sujet, disons, des quatre semaines qui ont précédé l'interview. On peut éliminer ce biais en demandant aux répondants de donner des renseignements au sujet d'une période de référence fixe.

Soulignons qu'il faut accorder plus d'attention au biais de non-contact dans le cas des enquêtes sur l'emploi du temps que dans celui d'autres enquêtes, car, contrairement

2. STRATÉGIES DE PRISE DE CONTACT, ACTIVITÉS CORRÉLÉES ET BIAIS D'ACTIVITÉ

À la présente section, nous comparons les biais d'activité associées au calendrier de prise de contact à journées convenables et à chacune des trois versions du calendrier à journées désignées. Ces calendriers sont définis comme suit :

1. Journées convenables (JC) : tentative de prise de contact avec les répondants potentiels chaque jour après la première tentative, jusqu'à ce qu'on joigne le répondant potentiel ou que la période de travail sur le terrain se termine.
2. Journées désignées (JD) : tentative unique de prise de contact avec les répondants (pas de tentatives subséquentes).
3. Journées désignées avec remise (JDR) : tentatives de prise de contact avec les répondants potentiels le même jour de la semaine que celui de la première tentative, jusqu'à ce qu'on joigne le répondant potentiel ou que la période de travail sur le terrain se termine.
4. Journées désignées avec remise et substitution (JDRS) : tentatives de prise de contact avec les répondants potentiels un jour sur deux après la première tentative jusqu'à ce qu'on joigne le répondant potentiel ou que la période de travail sur le terrain se termine.

La stratégie JDRS se fonde sur la supposition d'une alternance de tentatives de prise de contact les mardi/jeuvi et les mercredi/vencredi. Le fait que les journées de prise de

Le reste de l'article est présenté comme suit. À la section 2, nous décrivons quatre stratégies de prise de contact et nous utilisons des simulations simples pour évaluer le biais d'activité associé à chacune d'elles. À la section 3, nous étirons les simulations à l'aide de données provenant du Work Schedule Supplement to the Current Population Survey de mai 1997 et de la Time Diary Study réalisée en 1995 par l'Université du Maryland et nous examinons comment le biais varie selon le genre d'activité. En outre, nous décomposons le biais global, pour évaluer la contribution relative du biais d'activité et du biais de non-contact. Enfin, à la section 4, nous résumons les résultats et faisons des recommandations.

ne prennent jamais contact avec les répondants durant des journées PCD (c'est-à-dire que $P_D = 0$, où P_D est la probabilité d'un contact durant une journée PCD) et qu'ils prennent toujours contact avec les répondants durant des journées PCF (c'est-à-dire que $P_F = 1$, où P_F est la probabilité d'un contact durant une journée PCF). Enfin, supposons que la probabilité qu'une journée soit une journée PCF est de 0,5, si bien qu'en moyenne, la moitié des journées de chaque répondant potentiel sont PCF et l'autre moitié, PCD. Notons que les répondants potentiels sont tous identiques, en ce sens que la probabilité que toute journée soit une journée PCF est de 0,5 pour tous les répondants potentiels. Pour simplifier, nous supposons que les activités durant une journée particulière peuvent être résumées grâce à un « indice d'activité », I_j , où $I_j = 1 - P_j$ ($j = D, F$). L'indice d'activité représente le temps consacré aux activités qui sont corrélées négativement la probabilité d'un contact. Donc, les journées PCD sont des journées durant lesquelles plus de temps est consacré aux activités qui ont lieu en-dehors de la maison (travail, magasinage, loisirs, etc.), tandis que les journées PCF sont des journées durant lesquelles plus de temps est consacré aux activités entreprises à la maison (travaux ménagers, loisirs passifs, etc.). L'indice d'activité réelle moyenne pour la population de répondants possibles est égal à 0,5 ($= 0,5 \times 1 + 0,5 \times 0$).

Si l'on utilise un calendrier de prise de contact à journées convenables et que le nombre de rappels n'est pas limité, il y a suréchantillonnage des journées PCD. Pour comprendre pourquoi, examinons en détail les deux séries possibles de contacts. Si la première tentative de contact a lieu durant une journée PCF, il y a prise de contact avec le répondant auquel on demande de déclarer les activités de la journée précédente (la journée du journal). Comme la probabilité qu'il s'agisse d'une journée PCF ou PCD est la même, en moyenne, la moitié des journées du journal seront des journées PCD et l'autre moitié, des journées PCF. Par conséquent, l'indice d'activité moyen pour les journées du journal de ces répondants est égal à 0,5, c'est-à-dire la même valeur que pour la moyenne de la population. Si, par contre, la journée du premier contact est une journée PCD, aucune interview n'a lieu et on rappelle le répondant le jour suivant. Les tentatives de prise de contact se poursuivent chaque jour jusqu'à ce que le répondant potentiel soit rejoint (durant une journée PCF). Pour ce répondant, l'indice d'activité moyen pour les journées du journal est égal à 1, parce que le répondant est toujours interviewé durant une journée PCF qui suit directement une journée PCD. Donc, si une journée particulière est une journée PCD (c'est-à-dire durant laquelle le répondant potentiel entreprend beaucoup d'activités en-dehors de la maison), il est plus probable que cette journée soit sélectionnée comme journée de référence. Par conséquent, la probabilité d'interviewer le répondant potentiel durant une journée de référence particulière est corrélée aux activités qui ont lieu cette journée-là. Puisque la moitié des premières tentatives de

au biais introduit par la remise de l'interview, à l'hétérogénéité inobservée qui est corrélée à la probabilité de remise de l'interview ou simplement à un bruit aléatoire. Quoi qu'il en soit, ils soutiennent que les différences sont très faibles, si bien que tout biais éventuel est petit.

L'un des avantages du calendrier à journées convenables est qu'il est possible de procéder à de nombreuses tentatives de prise de contact en une brève période. En revanche, le calendrier à journées désignées – tel que proposé – ne permet de procéder qu'à une seule tentative par semaine. Donc, il est naturel de se demander s'il serait déraisonnable de modifier le calendrier à journées désignées pour permettre une certaine forme de substitution du jour de la semaine, par exemple, si le répondant ne peut être rejoint le mardi pour faire une déclaration concernant le lundi, serait-il acceptable de communiquer avec lui, disons, le jeudi en lui demandant de faire une déclaration concernant le mercredi? Ce calendrier modifié permettrait de faire un plus grand nombre de tentatives de prises de contact sans devoir prolonger la période de travail sur le terrain.

Comme ce genre de substitution n'est sensée que si les journées de remplacement sont assez semblables aux journées originales, la première étape consiste à déterminer quelles journées, si tant qu'il y en ait, sont similaires à d'autres. Lors de travaux antérieurs, Stewart (2000) nous a montré que, du lundi au jeudi, les journées sont fort semblables les unes aux autres, que les vendredis sont légèrement différents des jours de fin de semaine, et que le samedi et le dimanche sont très différents des jours de semaine et l'un de l'autre. Donc, il serait raisonnable de permettre des substitutions entre les jours de la semaine, au moins du lundi au jeudi.

Biais d'activité et biais de non-contact

Lors du choix d'une stratégie de prise de contact, nous devons tenir compte de deux types de biais, à savoir le biais d'activité et le biais de non-contact. Le biais d'activité survient lorsque la probabilité de joindre et d'interviewer un répondant potentiel durant une journée particulière est corrélée aux activités de ce répondant durant la journée correspondante pour laquelle il doit produire le journal de ses activités. Notons que dans la suite de l'article, l'expression probabilité d'un contact s'entend de la probabilité d'un contact productif (c'est-à-dire qui aboutit à une interview). Afin d'isoler les effets de diverses stratégies de prise de contact, on suppose que les répondants avec lesquels il y a eu contact acceptent systématiquement de participer à l'interview. Le biais de non-contact survient lorsque des différences d'activités font varier la probabilité d'un contact d'une personne à l'autre. Suivent des exemples numériques simples qui illustrent ces biais.

Exemple 1 – Biais d'activité : Supposons que les journées des répondants potentiels se répartissent en deux catégories, c'est-à-dire les journées à prise de contact difficile (journées PCD) et les journées à prise de contact facile (journées PCF). En outre, supposons que les intervieweurs

une déclaration concernant le dimanche). Les intervieweurs avaient pour instruction de faire au moins 20 tentatives d'appel avant de finaliser le cas comme étant non accompli. Les auteurs de la plupart des articles de méthodologie appuient l'utilisation du calendrier à journées désignées (Kinsley et O'Donnell 1983; Kalton 1985; Lyberg 1989; Harvey 1993 et Harvey 1999). Par exemple, Lyberg (1989) soutient que le calendrier à journées convenables peut introduire un biais, car « le répondant peut choisir une journée où il n'est pas occupé, une journée où il ne s'adonne pas à un comportement socialement inacceptable, une journée qu'il considère comme représentative, etc. ». Kinsley et O'Donnell (1983) soulignent, quant à eux, que l'horaire à journées convenables pourrait produire une exagération du nombre d'événements qui surviennent en-dehors de la maison, car le répondant est plus susceptible d'être interviewé durant une journée qui suit immédiatement une journée où il était hors de la maison. Deux de ces études font une comparaison directe des calendriers à journées désignées et à journées convenables (Kinsley et O'Donnell 1983; Lyberg 1989). Dans Kinsley et O'Donnell (1983), le plan d'expérience consiste à diviser l'échantillon en deux groupes. Ces auteurs ont observé que les deux catégories de calendriers produisaient des taux de réponse similaires et que la composition démographique était la même pour les deux échantillons. Ils ont également observé que le temps estimatif passé en-dehors de la maison était nettement plus important dans le cas du calendrier à journées convenables que dans celui du calendrier à journées désignées. Toutefois, il est impossible de déterminer si l'utilisation du calendrier à journées convenables donne lieu à une surestimation du temps passé en-dehors de la maison, ou si le calendrier à journées désignées produit une sous-estimation de ce temps, car on ne sait pas quelle est la vraie réponse. Dans Lyberg (1989), on a demandé au répondant de tenir deux journaux. La tenue du premier journal était basée sur un calendrier à journées désignées et l'autre, sur un calendrier à journées convenables. Cependant, le questionnaire du journal à journées convenables a été administré par l'intervieweur, tandis que celui du journal à journées désignées a été auto-administré par le répondant. Plusieurs jours après l'interview basée sur les journées convenables, il est impossible de déterminer si un écart donné est dû à des différences entre les calendriers de prise de contact ou à des effets de mode.

Deux études (Lyberg 1989; Laaksonen et Pääkkönen 1992) portent sur l'effet qu'a la remise de l'interview sur les taux de réponse. Dans les deux cas, les auteurs constatent que la remise de l'interview fait augmenter le taux de réponse. Laaksonen et Pääkkönen (1992) admettent aussi qu'il est difficile de déterminer si le fait de repousser l'interview introduit un biais. Leurs résultats montrent que les répondants qui font repousser l'interview consacrent moins de temps aux tâches ménagères et à l'entretien, et ne peuvent établir avec certitude si ces différences sont dues

à une date ultérieure et lui demander de déclarer ses activités lors de la journée originale de référence. Cette démarche permet de garder la journée de référence, mais allonge la période de remémoration. Cette démarche consiste à remettre l'interview et à attribuer au répondant une nouvelle journée de référence. Kalton (1985) recommande de repousser l'interview d'une semaine exactement, de sorte que la nouvelle journée de référence corresponde au même jour de la semaine que la journée de référence originale. Ces démarches ne sont pas mutuellement exclusives. Par exemple, le calendrier à journées désignées de Statistique Canada permet aux intervieweurs d'appeler les répondants jusqu'à deux jours après la journée de référence (Statistique Canada 1999) et de repousser l'interview d'une semaine si le répondant ne peut être rejoint après la deuxième journée d'essai. L'interview ne peut pas être remise de plus de trois semaines (Statistique Canada). Par exemple, si la journée de référence initiale est le lundi 1^{er}, on appelle le répondant le mardi 2^e, au besoin, le mercredi 3. Si aucune interview n'a lieu d'une de ces journées, on appelle le répondant le mardi 9^e, au besoin, le mercredi 10^e, et on lui demande de déclarer ses activités du lundi 8. Ce processus se poursuit jusqu'à ce que le répondant soit interviewé, qu'il refuse de participer ou que quatre semaines se soient écoulées.

Le calendrier à journées convenables pour le répondant ne permet pas de garder la journée de référence désignée. Si aucun contact n'est pris, l'intervieweur appelle la journée suivante et chaque journée subséquente jusqu'à ce qu'il rejoigne le répondant. Lorsqu'il y a eu contact, l'intervieweur essaye de procéder à l'interview ou, si le répondant refuse de participer à l'interview à ce moment-là, fixe une autre journée convenable pour le répondant. La journée de référence est toujours la journée qui a précédé l'interview. Souignons que puisque les répondants ne planifient vraisemblablement pas d'interview lors de journées occupées, leur permettre de choisir la journée de l'interview revient pratiquement au même que laisser à l'intervieweur la possibilité de proposer des journées consécutives (ou d'appeler pendant plusieurs journées consécutives) jusqu'à ce que le répondant potentiel accepte de répondre. Donc, on pourrait considérer le calendrier à journées convenables comme étant fonctionnellement identique à un calendrier avec tentatives de prise de contact chaque jour.

Une variante du calendrier à journées convenables décrit plus haut a été utilisée lors de l'Environnemental Protection Agency (EPA) Time Diary Study de 1992-1994, réalisée par l'Université du Maryland (voir Triplet 1995). Plutôt que d'affecter les répondants à une journée d'appel initiale, on les a répartis en un échantillon des jours de semaine et un échantillon des fins de semaine. Par exemple, les répondants faisant partie de l'échantillon des fins de semaine pouvaient être appelés le dimanche (pour faire une déclaration concernant le samedi) ou le lundi (pour faire

Evaluation du biais lié à diverses stratégies de prise de contact dans les enquêtes téléphoniques sur l'emploi du temps

JAY STEWART¹

RÉSUMÉ

Dans le cas de la plupart des enquêtes téléphoniques sur l'emploi du temps, on appelle les répondants potentiels une journée donnée pour leur demander de déclarer leurs activités de la journée précédente. Comme la plupart ne sont pas disponibles la journée du premier appel, cette méthode d'enquête fait que la probabilité d'interviewer la personne au sujet d'une journée de référence particulière risque d'être corrélée aux activités qui ont eu lieu durant cette journée de référence. De surcroît, le biais de non-contact a plus d'importance dans le cas des enquêtes sur l'emploi du temps que dans les autres, parce que les réponses par procuration ne peuvent être acceptées. Par conséquent, il est essentiel, pour ces enquêtes, d'adopter une stratégie prévoyant des tentatives subséquentes de prise de contact avec les répondants. Une stratégie de prise de contact spécifique le calendrier des prises de contacts et la période de travail sur le terrain. Les auteurs de publications antérieures ont défini deux calendriers pour procéder à ces tentatives subséquentes : un calendrier basé sur les journées convenables pour le répondant et un calendrier basé sur des journées désignées. La plupart de ces auteurs recommandent d'adopter le calendrier à journées désignées, mais offrent peu de données pour étayer ce choix. Dans le présent article, nous utilisons des simulations informatiques pour examiner le biais associé au calendrier à journées convenables et à trois versions du calendrier à journées désignées. Les résultats indiquent qu'il est préférable d'utiliser le calendrier à journées convenables et avant tout, les estimations fondées sur le calendrier à journées désignées sont sensibles à la variance de la probabilité d'un contact. Par contre, un calendrier à journées désignées avec possibilité de remise de l'interview ne crée qu'un biais très faible et produit des résultats robustes pour un large éventail d'hypothèses quant au profil des activités durant les diverses journées de la semaine.

MOTS CLÉS : Enquêtes téléphoniques sur l'emploi du temps; stratégies de prise de contact; biais; simulations informatiques.

1. INTRODUCTION

Les enquêtes téléphoniques sur l'emploi du temps posent un problème unique de collecte des données, parce qu'on appelle les répondants une journée particulière pour leur demander de déclarer leurs activités de la journée précédente. La difficulté tient au fait qu'on n'arrive pas à joindre la plupart des répondants – environ 75 % (Kallion 1985) – lors du premier appel, ce qui oblige à faire d'autres tentatives de prise de contact. Dans le cas de la plupart des enquêtes, le moment où ces tentatives supplémentaires sont faites n'importe pas, car on demande aux répondants de faire une déclaration concernant une période de référence fixe. En outre, dans la plupart des enquêtes, le problème de remémoration n'est pas trop important même si on reprend contact avec les répondants plusieurs jours après le premier appel. Par contre, dans le cas des enquêtes sur l'emploi du temps, la capacité qu'ont les répondants de se souvenir de leurs activités durant une journée particulière diminue considérablement après un jour ou deux, si bien qu'il faut attribuer une nouvelle journée de référence au répondant si on ne réussit pas à le joindre au moment du premier appel. Comme nous le verrons plus loin, dans de telles conditions, il se peut que la probabilité d'interviewer le répondant au sujet d'une journée de référence particulière soit reliée aux

activités qui ont eu lieu durant cette journée de référence. Par conséquent, il est essentiel, pour ce genre d'enquête, d'étayer une stratégie de prise subséquentes de contact avec le répondant qui n'introduit pas de biais.

Stratégies de prise de contact

Une stratégie de prise de contact comprend un calendrier de prise de contact et une période de travail sur le terrain. Le calendrier de prise de contact spécifie quelles journées de la semaine les tentatives de contact seront faites et la période de travail sur le terrain précise le nombre maximum de semaines durant lesquelles des tentatives seront faites. Les calendriers de prise de contact rentrent dans deux grandes catégories : les calendriers à journées désignées et les calendriers à journées convenables. Pour l'une et l'autre de ces catégories, on affecte aléatoirement chaque répondant à une journée initiale d'appel. S'il y a effective-ment prise de contact lors de ce premier appel, l'intervieweur essaye de recueillir l'information au sujet de la journée de référence, qui est la journée précédant celle de l'appel. Par contre, à l'étape des tentatives subséquentes de prise de contact, les deux stratégies diffèrent. Dans le cas d'un calendrier à journées désignées, deux démarches peuvent être adoptées pour faire les tentatives subséquentes de prise de contact. L'intervieweur pourrait

- NANDRAM, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.
- NATIONAL CENTER FOR HEALTH STATISTICS (1992). Third National Health and Nutrition Examination Survey. *Vital and Health Statistics Series*, 2, 113.
- NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey. *Vital and Health Statistics, Series* 1, 32.
- OLSON, R.L. (1980). A least squares correction for selectivity bias. *Econometrica*, 48, 1815-1820.
- PARK, T. (1998). An approach to categorical data nonignorable nonresponse. *Biometrics*, 54, 1579-1590.
- PARK, T., et BROWN, M.B. (1994). Models for categorical data with nonignorable non-response. *Journal of the American Statistical Association*, 89, 44-52.
- RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons, Inc.
- RUBIN, D.B. (1976). Inference et missing data. *Biometrika*, 63, 581-590.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- SCHAFER, J.L., EZZATI-RICE, T.M., JOHNSON, W., KHARE, M., LITTLE, R.J.A. et RUBIN, D.B. (1996). The NHANES III multiple imputation project. Survey Research Methods, *Proceedings of the American Statistical Association*, 28-37.
- STASNY, E.A. (1991). Hierarchical models for the probabilities of a survey classification et nonresponse: An example from the national crime survey. *Journal of the American Statistical Association*, 86, 296-303.
- STASNY, E.A., KADANE, J.B. et FRITSCH, K.S. (1998). On the fairness of death penalty jurors: A comparison of bayesian models with different levels of hierarchy and various missing-data mechanisms. *Journal of the American Statistical Association*, 93, 464-477.

Nous avons exécuté l'échantilleur MH en établissant un écart aléatoire à partir de chacune des valeurs de (A.3), (A.4) et (A.5). Il est facile de prélever un écart aléatoire à partir de (A.5). Nous avons obtenu des échantillons de chacune des valeurs de (A.3), (A.4) et (A.5) au moyen de l'algorithme de Nandram (1998).

ANNEXE 2

Estimations de la régression non linéaire selon la méthode des moindres carrés

Soit

$$v_{j1l} = \log \left\{ \sum_{i=1}^s q_{isl} / \left(1 - \sum_{j=1}^s q_{isl} \right) \right\}, j = 1, \dots, J-1 = J'.$$

Les valeurs de v_{j1l} sont obtenues pour chaque itération d'après l'échantilleur Metropolis-Hastings. Pour régler le problème que pose la régression non linéaire selon la méthode des moindres carrés, nous avons minimisé

$$(A.1) \quad \left\{ \sum_{j=1}^J \sum_{l=1}^L \left\{ v_{j1l} - e^{\phi_l} (\theta_j - (\mu_l + \alpha_l)) \right\}^2 \right\}$$

sous réserve des contraintes $\sum_{l=1}^L \mu_l = 0$, $\sum_{l=1}^L \alpha_l = 0$, $\sum_{l=1}^L \mu_l = 0$, et soit $e^{\phi_l} = \psi_l$, $\sum_{l=1}^L \psi_l = 1$ in $\psi_l = 0$.

En s'appuyant sur les dérivées partielles dans le but d'obtenir l'estimation des moindres carrés, nous avons

$$(A.2) \quad \left\{ \frac{\sum_{j=1}^J \sum_{l=1}^L v_{j1l} (\theta_j - \mu_l - \alpha_l)}{\sum_{j=1}^J \sum_{l=1}^L \left(\theta_j - \mu_l - \alpha_l \right)^2} \right\} = \log \psi_l^{-1}$$

où

$$\theta_j = \left\{ \sum_{l=1}^L e^{2\phi_l} \left\{ \frac{1}{8} \sum_{i=1}^8 \left(e^{-\phi_l} v_{j1l} + \mu_l + \alpha_l \right) \right\} \right\} / \left\{ \sum_{l=1}^L e^{2\phi_l} \right\}, \quad (A.3)$$

$$\mu_l = \left(\frac{1}{8} \sum_{j=1}^J \sum_{l=1}^L \left\{ \theta_j - \left(\alpha_l + e^{-\phi_l} v_{j1l} \right) \right\} \right) \quad (A.4)$$

et

$$\alpha_l = \left\{ \sum_{j=1}^J \frac{1}{J'} \sum_{l=1}^L e^{2\phi_l} \left\{ \theta_j - (\mu_l + e^{-\phi_l} v_{j1l}) \right\} \right\} / \left\{ \sum_{l=1}^L e^{2\phi_l} \right\}. \quad (A.5)$$

À partir de ces paramètres, nous avons prélevé les valeurs de q_{j1l} selon un algorithme MH. En outre, nous réglons le problème de la régression non linéaire selon la méthode des moindres carrés en appliquant la méthode itérative qui nous permet d'obtenir les valeurs de ϕ_l , θ_j , μ_l et α_l . Soit

$$v_{j1l}^{(h)} = \log \left\{ \sum_{i=1}^s q_{isl}^{(h)} / \left(1 - \sum_{j=1}^s q_{isl}^{(h)} \right) \right\},$$

BIBLIOGRAPHIE

- ou $q_{isl}^{(h)}$ dénote la valeur $q_{isl}^{(h)}$ à la $h^{\text{ème}}$ itération de l'algorithme MH. Puis, nous minimisons (A.1) sous réserve des contraintes susmentionnées à la $h^{\text{ème}}$ itération pour obtenir $\phi_l^{(h)}$, $\theta_j^{(h)}$, $\mu_l^{(h)}$ et $\alpha_l^{(h)}$. Ces itérations nous donnent une estimation des répartitions a posteriori de $\phi_l^{(h)}$, $\mu_l^{(h)}$ et $\alpha_l^{(h)}$. Il y a convergence dans le cas de notre application dans moins de 10 itérations.
- ALBERT, J.H., et GUPTA, A.K. (1985). Bayesian methods for binomial data with applications to a nonresponse problem. *Journal of the American statistical Association*, 80, 167-174.
- BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American statistical Association*, 83, 62-69.
- BASU, D., et PEREIRA, C.A. (1982). On the Bayesian analysis of categorical data: The Problem of nonresponse. *Journal of Statistical Planning et Inference*, 6, 345-362.
- CRAWFORD, S.L., JOHNSON, W.G. et LAIRD, N.M. (1993). Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey (avec discussion). Dans *Case Studies in Bayesian Statistics* (Eds. C. Gatsonis, J.S. Hodges, R.E. Kass et N.D. Singpurwalla). New York: Springer-Verlag, 78-117.
- DE HEER, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, 15, 129-142.
- DEBEY, J.J., et LINDLEY, D.V. (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association*, 76, 833-841.
- FORSTER, J.J., et SMITH, W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society, Series B*, 60, 57-70.
- GROVES, R.M., et COOPER, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection et limited dependent variables et a simple estimator for such models. *Annals of Economic et Social Measurement*, 5, 475-492.
- KUCZMARSKI, R.J., CARROL, M.D., FLEGAL, K.M. et TROIANO, R.P. (1997). Varying body mass index cutoff points to describe overweight prevalence among U. S. adults: NHANES III (1988 to 1994). *Obesity Research*, 5, 542-548.
- KAUFMAN, G.M., et KING, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association*, 68, 670-678.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- MALTEC, D., DAVIS, W. et CAO, X. (1999). Model-based small area estimates of over-weight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.
- MOHADIJER, L., BELL, B. et WAKSBERG, J. (1994). National Health and Nutrition Examination Survey III-accounting for item nonresponse bias. *National Center for Health Statistics*.

Dans le tableau I 1, nous montrons que tous les intervalles crédibles renferment les valeurs réelles et que les moyennes a posteriori se rapprochent de la valeur réelle dont l'écart le moins important concerne les cas de non-réponse ignorable. Les écarts-types sont très semblables dans les neuf exemples faibles et semblables dans les neuf exemples simulés. De plus, les erreurs-types numériques (é-t.n.) sont faibles et semblables dans les neuf exemples simulés. Les estimations de $\Pr(\delta < \delta^{(b)} | y, r)$ varient de 0,30 à 0,40, sauf pour les cas de non-réponse dont il faut le plus tenir compte pour lesquels $(a, b) = (0,8, 0,8)$ et $(0,8, 0,9)$. Ainsi, le modèle donne d'assez bons résultats.

6. CONCLUSION

Nous avons décrit une méthode bayésienne qui permet d'analyser les données multinomiales dans le cas des petites régions en l'absence de non-réponses non-ignorable. Nous avons employé un modèle hiérarchique qui donne d'assez bons résultats. En fait, nous avons élargi le champ d'application de la méthode de Stasny (1991) dans deux directions : a) nous avons étudié les données multinomiales renfermant plus de deux cellules (binomiales) et b) nous avons effectué une analyse bayésienne complète. Les points a) et b) ont été appliqués aux petites régions.

La méthode de Monte Carlo à chaînes de Markov nous a permis d'évaluer la structure complexe de l'estimation de la non-réponse multinomiale. D'après notre analyse empirique et notre étude de simulation, le modèle appliqué à ces données donne de bons résultats. Par conséquent, la méthode de l'estimation par ratio utilisée à l'heure actuelle dans la NHANES III peut être remplacée par notre méthode bayésienne vu que les caractéristiques des non-répondants peuvent différer de celles des répondants. En fait, l'application de notre modèle aux données de la NHANES III montre que dans chaque comté, il y a des différences importantes dans les proportions de personnes aux trois niveaux de l'IMC selon l'âge et le sexe. Nous pouvons l'observer au tableau I où sont additionnées les dénombrements des comtés. Cependant, nous avons obtenu une inférence (y compris la mesure de précision) pour chaque comté selon l'âge, la race et le sexe.

Nous pouvons élargir le champ d'application de notre méthode de trois façons. D'abord, nous pouvons utiliser un modèle qui incorpore une mesure du degré non-ignorable de la non-réponse plutôt que seulement la dichotomie entre non-réponse ignorable et non-réponse non-ignorable. Puis, nous pouvons utiliser d'autres répartitions a priori (par exemple, le processus a priori de Dirichlet) pour modéliser l'hétérogénéité de la mise en grappes des régions plutôt que de presumer l'homogénéité des régions comme nous l'avons fait. Par ailleurs, on peut ajouter une quatrième étape à notre modèle de manière à tenir compte de la mise en grappes au sein des ménages, de même que la mise en grappes au sein des régions (comtés) dans la NHANES III. Il s'agit là de tâches très difficiles.

Echantilleurs Metropolits-Hastings

Dans le cas du modèle de non-réponse ignorable, (μ_1, τ_1) et (μ_{21}, τ_{21}) sont des valeurs indépendantes a posteriori à

(A.1)
$$p(\mu_1, \tau_1 | y, r) \propto p(\mu_1, \tau_1) \prod_{i=1}^t \left\{ \frac{D(y_i + n_i - r_i + \mu_1 \tau_1)}{D(y_i + \mu_1 \tau_1)} \right\}$$

et

(A.2)
$$p(\mu_{21}, \tau_{21} | y, r) \propto p(\mu_{21}, \tau_{21}) \left\{ \frac{\prod_{i=1}^t B(r_i + \mu_{21} \tau_{21}, r_i - y_i + (1 - \mu_{21}) \tau_{21})}{B(\mu_{21} \tau_{21}, (1 - \mu_{21}) \tau_{21})} \right\}$$

où $p(\mu_1, \tau_1)$ et $p(\mu_{21}, \tau_{21})$ constituent les répartitions a priori. On peut obtenir des échantillons de (A.1) et (A.2) au moyen de l'algorithme MH de Nandram (1998).

Dans le cas du modèle de non-réponse non-ignorable, il est approprié de s'appuyer sur z pour obtenir

(A.3)
$$p(\mu_3, \tau_3 | z, y, r) \propto p(\mu_3, \tau_3) \prod_{j=1}^t \left\{ \frac{D(y_j + z_j + \mu_3 \tau_3)}{D(\mu_3 \tau_3)} \right\}$$

où $p(\mu_3, \tau_3)$ et $p(\mu_{4j}, \tau_{4j})$ sont des valeurs indépendantes avec

(A.4)
$$p(\mu_{4j}, \tau_{4j} | z, y, r) \propto p(\mu_{4j}, \tau_{4j}) \left\{ \frac{B(\mu_{4j} \tau_{4j}, (1 - \mu_{4j}) \tau_{4j})}{B(y_j + \mu_{4j} \tau_{4j}, z_j + (1 - \mu_{4j}) \tau_{4j})} \right\},$$

(A.5)
$$p(z_{11} = t_{11}, \dots, z_{tJ} = t_{tJ} | y, r, \mu_4, \tau_4, \mu_{4j}, \tau_{4j}, j = 1, \dots, J) =$$

$$\frac{\sum_{n_j=0}^{t_j} \dots \sum_{n_t=0}^{t_t} W^{h_{11}t_{11}, \dots, h_{tJ}t_{tJ}}}{\sum_{n_j=0}^{t_j} \dots \sum_{n_t=0}^{t_t} W^{h_{11}t_{11}, \dots, h_{tJ}t_{tJ}}}$$

pour $t_j = 0, 1, \dots, n_j - r_j$ $t_j = n_j - r_j$

$$W^{h_{11}t_{11}, \dots, h_{tJ}t_{tJ}} = \binom{n_1 - r_1}{t_{11}} \dots \binom{n_t - r_t}{t_{tJ}} \left(D(y_1 + t_1 + \mu_4 \tau_4) \right)$$

$$\prod_{j=1}^J B(y_j + \mu_{4j} \tau_{4j}, t_{4j} + (1 - \mu_{4j}) \tau_{4j}).$$

REMERCIEMENTS

Les travaux ont été effectués à la National Center For Health Statistics. Balgobin Nandram était le premier chercheur universitaire ASA/NCHS et Gounshik Han était en congé sabbatique de l'Université Hanshin, Corée.

ANNEXE I

Tableau 9
Comparaison des intervalles crédibles à 95 % pour θ_1, θ_2 et $\alpha_1, \dots, \alpha_8$ pour les groupes de personnes les plus jeunes et les plus âgées selon le type de régression

Non-linéaire	
θ_1	(-1,743, -1,469)
θ_2	(0,028, 0,196)
α_1	(-1,167, -0,751)
α_2	(-1,395, -0,939)
α_3	(-1,127, -0,723)
α_4	(-1,112, -0,659)
α_5	(1,198, 1,514)
α_6	(0,513, 0,689)
α_7	(0,715, 1,210)
α_8	(0,809, 1,310)

Tableau 10
Comparaison des intervalles crédibles à 95 % pour θ $\alpha_1, \dots, \alpha_4$ concernant le groupe des personnes les plus jeunes selon le type de régression

Linéaire	
θ	(1,455, 1,729)
α_1	(0,165, 0,592)
α_2	(-0,535, 0,014)
α_3	(0,078, 0,346)
α_4	(-0,704, -0,165)

5. UNE ÉTUDE DE SIMULATION

Nous décrivons une petite étude de simulation qui vise à évaluer le rendement de notre modèle multinomial de non-réponse non-ignorable. Nous mettons l'accent sur la probabilité de réponse.

Nous utilisons les données observées auprès des jeunes hommes blancs afin d'obtenir les moyennes à posteriori de P_{11}, P_{12}, P_{13} et $\pi_{11}, \pi_{12}, \pi_{13}$ pour chaque comté. Nous les considérons comme étant les valeurs vraies (i) que nous dénotons par $P_{11}^{(i)}, P_{12}^{(i)}, P_{13}^{(i)}$ et $\pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{13}^{(i)}$. Ainsi, la vraie probabilité de réponse dans le $i^{\text{ème}}$ comté est

$$\delta^{(i)} = \sum_{j=1}^J P_{ij}^{(i)} \pi_{ij}^{(i)} / \sum_{j=1}^J \pi_{ij}^{(i)} \quad \text{et} \quad \delta^{(i)} = \sum_{j=1}^J \delta_j^{(i)} / \sum_{j=1}^J \pi_j^{(i)}.$$

Dans nos exemples simulés, nous avons utilisé les n_i comme pour les données sur l'IMC qui concernent les jeunes hommes blancs et nous avons conservé les $P_{ij}^{(i)}$ fixes. Cependant, nous avons varié les $\pi_{ij}^{(i)}$ de la manière suivante. Nous avons conservé π_{11} fixe à $\pi_{11}^{(i)}$ et dénoté le vecteur des π_{1i} par π_1 . Les 34 valeurs des π_{1i} varient de 0,73 à 0,83. Puis, nous avons établi $\pi_2 = \alpha \pi_1$ et $\pi_3 = b \pi_1$, où $a, b = 0,8, 0,9, 1,0$. (Nous dénotons les vecteurs des π_{1i} et des π_{1i} par π_1 et π_2 respectivement.) Ainsi, il y a neuf exemples simulés.

Alors, pour chaque (a, b), nous avons produit des comptes pour une fonction de masse de probabilité

Note : moy. = moyenne à posteriori; $\hat{\epsilon}$ -t = erreur-type; $\hat{\epsilon}$ -1-n. = erreur-type numérique; IC = intervalle crédible à 95 %; prob. = $\Pr(\delta < \delta^{(i)} | y, r)$; les 34 valeurs de π_1 varient de 0,73 à 0,83.

π_2		$0,8 * \pi_1$		$0,9 * \pi_1$		$1,0 * \pi_1$	
stat	$0,8 * \pi_1$	stat	$0,8 * \pi_1$	stat	$0,9 * \pi_1$	stat	$1,0 * \pi_1$
prob.	0,82	prob.	0,82	prob.	0,82	prob.	0,82
IC	(0,678, 0,742)	IC	(0,678, 0,742)	IC	(0,678, 0,742)	IC	(0,678, 0,742)
$\hat{\epsilon}$ -1-n.	30	$\hat{\epsilon}$ -1-n.	30	$\hat{\epsilon}$ -1-n.	30	$\hat{\epsilon}$ -1-n.	30
moy.	712	moy.	712	moy.	712	moy.	712
réel	690	réel	690	réel	719	réel	719
prob.	0,37	prob.	0,37	prob.	0,37	prob.	0,37
IC	(0,673, 0,742)	IC	(0,673, 0,742)	IC	(0,673, 0,742)	IC	(0,673, 0,742)
$\hat{\epsilon}$ -1-n.	3	$\hat{\epsilon}$ -1-n.	3	$\hat{\epsilon}$ -1-n.	3	$\hat{\epsilon}$ -1-n.	3
moy.	726	moy.	726	moy.	726	moy.	726
réel	722	réel	722	réel	751	réel	751
prob.	0,303	prob.	0,303	prob.	0,303	prob.	0,303
IC	(0,712, 0,769)	IC	(0,712, 0,769)	IC	(0,712, 0,769)	IC	(0,712, 0,769)
$\hat{\epsilon}$ -1-n.	16	$\hat{\epsilon}$ -1-n.	16	$\hat{\epsilon}$ -1-n.	16	$\hat{\epsilon}$ -1-n.	16
moy.	742	moy.	742	moy.	742	moy.	742
réel	735	réel	735	réel	764	réel	764
prob.	0,95	prob.	0,95	prob.	0,95	prob.	0,95
IC	(0,708, 0,767)	IC	(0,708, 0,767)	IC	(0,708, 0,767)	IC	(0,708, 0,767)
$\hat{\epsilon}$ -1-n.	15	$\hat{\epsilon}$ -1-n.	15	$\hat{\epsilon}$ -1-n.	15	$\hat{\epsilon}$ -1-n.	15
moy.	758	moy.	758	moy.	758	moy.	758
réel	751	réel	751	réel	784	réel	784
prob.	0,399	prob.	0,399	prob.	0,399	prob.	0,399
IC	(0,693, 0,757)	IC	(0,693, 0,757)	IC	(0,693, 0,757)	IC	(0,693, 0,757)
$\hat{\epsilon}$ -1-n.	36	$\hat{\epsilon}$ -1-n.	36	$\hat{\epsilon}$ -1-n.	36	$\hat{\epsilon}$ -1-n.	36
moy.	784	moy.	784	moy.	784	moy.	784
réel	778	réel	778	réel	809	réel	809
prob.	0,318	prob.	0,318	prob.	0,318	prob.	0,318
IC	(0,725, 0,784)	IC	(0,725, 0,784)	IC	(0,725, 0,784)	IC	(0,725, 0,784)
$\hat{\epsilon}$ -1-n.	26	$\hat{\epsilon}$ -1-n.	26	$\hat{\epsilon}$ -1-n.	26	$\hat{\epsilon}$ -1-n.	26
moy.	809	moy.	809	moy.	809	moy.	809
réel	809	réel	809	réel	809	réel	809

Tableau 11
Caractéristiques de la probabilité de réponse

multinomiale avec les probabilités $P_{11}^{(i)}, P_{12}^{(i)}, P_{13}^{(i)}$ ($1 - \pi_{11}$). Nous dénotons ces dénombrements de cellules par $y_{11}, y_{12}, y_{13}, z_{11}, z_{12}, z_{13}$ et le nombre de répondants est $r_i = \sum_{j=1}^J y_{ij}$. Par la suite, nous ajustons le modèle de non-réponse non-ignorable selon les données ci-dessus au moyen de l'échantillonneur MH et nous obtenons les valeurs $M = 1\ 000$ ($P_{11}^{(h)}, \pi_{11}^{(h)}, h = 1, \dots, M$. Pour chaque valeur, nous avons calculé $\hat{\delta}^{(h)} = \sum_{j=1}^J \delta_j^{(h)} / \sum_{j=1}^J \pi_j^{(h)}$ ou $\hat{\delta}^{(h)} = \sum_{j=1}^J P_{ij}^{(h)} \pi_{ij}^{(h)} / \sum_{j=1}^J \pi_{ij}^{(h)}$. Au tableau 11, nous consignons les moyennes à posteriori, les écarts-types, les erreurs-types numériques (au moyen de la méthode des moyennes par lot) et l'intervalle crédible à 95 % pour la probabilité de réponse pour chaque choix de (a, b). Nous avons également calculé $\Pr(\hat{\delta} < \delta^{(i)} | y, r)$ en comptant le nombre de $\delta^{(i)}$ qui sont aussi importants que $\delta^{(i)}$. Il y a défaillance du modèle si cette valeur est extrêmement élevée ou faible.

Nous avons représenté graphiquement les estimations des densités à posteriori de δ par les choix de a et b que nous avons obtenus au moyen de l'estimateur de densité kernel normal avec une fenêtre optimale d'après l'analyse des sorties de l'algorithme MH. Les densités sont unimodales, avec pointe et presque symétriques. En augmentant (a, b), c'est-à-dire en la faisant passer de (0,8,0,8) à (1,0, 1,0), le mode des densités à posteriori augmente.

intervals pour d_j

Comparaison des modèles de non-réponse ignorable (ig) et non-ignorable (mig) pour les quatre exemples (Ex) correspondant aux petits domaines qui s'appuient sur les probabilités de cellules (d^f) et la probabilité de réponse (d^r)

	1	2
ig	0.444	0.480
mo	0.308	0.480
et	0.073	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.067	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444
et	0.067	0.079
ic	0.073	0.082
mo	0.444	0.459
ig	0.248	0.133
mo	0.308	0.448
et	0.067	0.073
ic	0.073	0.273
mo	0.444	0.450
ig	0.248	0.125
mo	0.308	0.386
et	0.067	0.093
ic	0.073	0.256
mo	0.444	0.638
ig	0.248	0.137
mo	0.308	0.444

	4	ig	mg
mon.	0.448		
6-1	0.089		
6-1	0.075		
mon.	0.265		
mon.	0.116, 0.591(0.186, 0.483)(0.607, 0.836)		
mon.		0.288	
6-1		0.081	
mon.		0.430	
6-1		0.10	
6-1		0.217, 0.619(0.104, 0.453)(0.145, 0.517)(0.768, 0.948)	

Enfin, dans le tableau 8, nous étudions les estimations de p_{ij} pour les comités individuels dont la moyenne globale \bar{p}_j figure dans le tableau 7. La situation est sensiblement la même que pour les p_{ij} . Cependant, nous observons que le premier exemple constitue une exception où l'intervalle crédible pour δ (0,459, 0,773) est presque entièrement du côté gauche de l'intervalle crédible pour δ (0,735, 0,801). Ainsi, on constate un retournement considérable attribuable au nombre relativement important de non-répondants, 14 dans le comité à l'égard des hommes blancs de 45-

et non linéaire

Disons que $q_{fjl}^{l^{\text{ieme}}}$ dénote la probabilité qu'un répondant du l^{ieme} groupe âge-race-sexe dans le j^{ieme} niveau de l'IMC. (Nous ajoutons l'indice inférieur l aux d_{fj} pour des raisons techniques). Soit $v_{fjl}^{l^{\text{ieme}}} = \log \{ \sum_{j=1}^J d_{fjl}^{l^{\text{ieme}}} / \sum_{j=1}^J d_{fj}^{l^{\text{ieme}}} \}$, $l = 1, \dots, J-1$, $j = 1, \dots, J$, nous prenons

$${}^1\mathfrak{h} / (({}^1\mathfrak{x} + {}^1\mathfrak{n}) - {}^f\theta) = {}^f\mathfrak{h}_\Lambda$$
$$(1-f) \circ 8^{1/f} \circ \frac{1}{8} \circ \frac{1}{1-f} \circ \frac{1}{8} = \dots \circ \frac{1}{8}$$
$$(1-f)8^{1/f} \Delta = \sum_{l=1}^f \Delta = f \Delta$$

grand et à la droite de celui qui concerne le modèle linéaire.

Tableau 6
Les estimations ponctuelles et les intervalles crédibles à 95 % pour la probabilité de réponse pondérée, $\hat{\delta} = \sum_{i=1}^n n_i p_{ij} / \sum_{i=1}^n n_i$, à l'égard des trois choix de Ω et du groupe des personnes les plus jeunes.

Race Sexe	$\hat{\delta}$	$\delta - t(\hat{\delta})$	Ω		$\Omega/4$	
			Intervalle	$\delta - t(\hat{\delta})$	Intervalle	$\delta - t(\hat{\delta})$
B	H 0,775	(0,744, 0,805)	0,769	0,017	0,735, 0,801	0,767
	F 0,855	(0,821, 0,886)	0,855	0,020	(0,810, 0,887)	0,853
N	H 0,786	(0,752, 0,817)	0,780	0,018	(0,740, 0,813)	0,778
	F 0,880	(0,854, 0,902)	0,878	0,015	(0,845, 0,903)	0,876
						0,015
						(0,838, 0,903)

Nota : Voir la note au tableau 1.

Au tableau 6, on trouve les estimations ponctuelles de la probabilité de réponse $\hat{\delta}$, et leurs intervalles crédibles à 95 % pour trois choix de Ω . Les probabilités de réponse pour les hommes sont inférieures à celles des femmes, et cette tendance demeure la même pour les trois choix de Ω . Si une enquête semblable est menée dans le futur, il faudrait que nous augmentions la taille de l'échantillon de 1,30 = (1/0,769) fois pour les hommes blancs et 1,17 = (1/0,855) fois pour les femmes blanches (c'est-à-dire que si l'intervaleur doit recueillir des données complètes auprès de 1 000 ménages, il devra communiquer avec 1 300 hommes blancs).

Au tableau 7, nous présentons les intervalles crédibles à 95 % pour les p_j pour les trois niveaux de l'IMC. Pour ce qui est des personnes plus jeunes, p_1 du niveau 1 de l'IMC est le plus élevé, et p_2 du niveau 2 de l'IMC est le plus élevé, et p_3 du niveau 3 de l'IMC est le plus faible. En particulier, les valeurs de p_1 , p_2 sont élevées et la valeur de p_3 est faible pour les hommes blancs.

Comme l'a suggéré un examinateur, nous avons étudié en détail les résultats concernant les femmes blanches plus âgées (45+) au tableau 1, les proportions observées dans les trois niveaux de l'IMC sont 0,079, 0,347 et 0,568. Toutefois, les intervalles crédibles à 95 % qui s'appliquent aux proportions de la population au tableau 7 sont (0,059, 0,068), (0,431, 0,451) et (0,486, 0,505) respectivement. Autrement dit, alors que les proportions observées se rapprochent des intervalles, aucun de ces intervalles ne renferme les proportions observées.

Nous pouvons expliquer le phénomène de la façon suivante. Les données pour les femmes blanches plus âgées (45+) sont très claires. Pour les 34 comtés, les quartiles des comptes observés dans les trois niveaux de l'IMC sont (0,1,3), (3,6,10) et (5,9,14) respectivement. Par conséquent, quand le modèle de non-réponse ignorable est ajusté selon les 34 comtés, il y a réduction non seulement entre les comtés, mais aussi d'un niveau de l'IMC à l'autre. En conséquence, la proportion supérieure tend à être plus faible et la proportion la plus faible tend à être supérieure, et

Nota 1 : Le modèle de non-réponse non-ignorable est appliqué au groupe des personnes les plus jeunes.
Nota 2 : Le modèle de non-réponse ignorable est appliqué au groupe des personnes les plus âgées.

Âge Race Sexe	p_1	p_2	p_3	Intervalle crédible à 95 %	
				$p_j - t(p_j)$	$p_j + t(p_j)$
45-	H	(0,382, 0,470)	(0,174, 0,252)	(0,314, 0,412)	(0,443, 0,371)
	F	(0,425, 0,525)	(0,171, 0,269)	(0,333, 0,419)	(0,443, 0,710)
45+	H	(0,381, 0,445)	(0,176, 0,241)	(0,329, 0,442)	(0,486, 0,505)
	F	(0,202, 0,041)	(0,255, 0,326)	(0,040, 0,093)	(0,592, 0,670)
N	H	(0,059, 0,068)	(0,431, 0,451)	(0,035, 0,076)	(0,486, 0,505)
	F	(0,040, 0,035)	(0,206, 0,265)	(0,035, 0,076)	(0,661, 0,731)

À partir des quatre premiers exemples du tableau 2, nous illustrons les estimations régionales. Comme on peut l'imaginer, il est fastidieux de présenter toutes les estimations pour les 34 comtés et les huit domaines. Le tableau 8 montre les moyennes à posteriori, les écarts-types et les intervalles crédibles à 95 % pour les p_{ij} et les δ_j . D'abord, nous comparons les estimations des p_{ij} d'après le modèle de non-réponse non-ignorable et le modèle de non-réponse ignorable. En général, les estimations d'après les deux modèles diffèrent : les intervalles dans le cas du modèle de non-réponse non-ignorable sont plus grands que dans le cas du modèle de non-réponse ignorable.

4. UNE ANALYSE DES DONNÉES DE LA NHANES III

Dans la présente section, nous illustrons notre méthode à partir des données sur l'IMC tirées de la NHANES III. Nous étudions d'abord nos estimations en fonction de mesures sommatoires par rapport aux comtés. Plus précisément, nous utilisons les distributions pondérées à posteriori des P_{ij} ,

$$\hat{P}_j = \sum_{i=1}^I n_i P_{ij} / \sum_{i=1}^I n_i, j = 1, 2, 3$$

et la distribution pondérée à posteriori des δ_i

$$\hat{\delta} = \sum_{i=1}^I n_i \delta_i / \sum_{i=1}^I n_i$$

pour chacun des huit domaines âge-race-sexe. Puis, pour les quatre premiers exemples figurant dans le tableau 2, nous indiquons les effets sur les petites régions.

Nous indiquons également comment mettre en relation les P_{ijk} et les π_{ij} avec l'âge, la race et le sexe à partir de modèles de régression logistique linéaire et non linéaire.

4.1 Analyse des données

Nous commençons par effectuer une analyse de sensibilité dans le but d'évaluer les spécifications de $\eta^{(0)}$ et $\nu^{(0)}$. Nous avons comparé trois choix d'hyper paramètres $\Omega = (\eta^{(0)}, \nu^{(0)})$ pour vérifier la sensibilité de la spécification des hyper-paramètres sur l'inférence. Notre premier choix correspond à quatre fois Ω , c'est-à-dire $4\Omega = (4\eta^{(0)}, 4\nu^{(0)})$; notre deuxième choix est les hyper-paramètres comme tels, c'est-à-dire $\Omega = (\eta^{(0)}, \nu^{(0)})$; et notre troisième choix correspond à un quatrième de Ω , c'est-à-dire, $\Omega/4 = (\eta^{(0)}/4, \nu^{(0)}/4)$. Dans le tableau 4, on trouve les résultats de simulation qui s'appliquent à la sensibilité de l'inférence de \hat{P}_j pour le groupe des personnes plus jeunes (45-). Les estimations ponctuelles et les écarts-types des proportions se ressemblent énormément dans les trois options d'hyper-paramètres. De même, le tableau 5 montre les résultats de simulation pour \hat{P}_j qui s'appliquent au groupe des personnes plus âgées (45+). Les estimations ponctuelles pour les hommes sont très semblables d'un choix à l'autre des hyper-paramètres. Dans le cas des femmes, nous observons toutefois de faibles variations dans les estimations ponctuelles, qui passent de 452 à Ω . Les écarts-types sont accrus lorsque Ω diminue pour les femmes; nous ne relevons toutefois aucune variation substantielle pour les hommes. En général, le modèle de non-réponse non-ignorable fonctionne mieux que le modèle de non-réponse ignorable, puisque le premier modèle n'est pas sensible au choix des hyper-paramètres.

Tableau 4
Sensibilité de \hat{P}_j pour le choix de $\eta_j^{(0)}, \nu_j^{(0)}$ et $\eta_{4j}^{(0)}, \nu_{4j}^{(0)}$, $j = 1, \dots, 4$ pour le groupe des personnes les plus jeunes (45-) pour les trois niveaux de l'IMC

Race	Sexe	\hat{P}_1	$\hat{P}_1 - \hat{P}_1^{(1)}$	\hat{P}_2	$\hat{P}_2 - \hat{P}_2^{(1)}$	\hat{P}_3	$\hat{P}_3 - \hat{P}_3^{(1)}$
------	------	-------------	-------------------------------	-------------	-------------------------------	-------------	-------------------------------

(a) 42	B	H	0.428	0.022	0.216	0.019	0.356	0.022
	F	0.476	0.025	0.232	0.020	0.292	0.024	0.020
(b) Ω	H	0.419	0.020	0.212	0.016	0.369	0.020	0.027
	F	0.434	0.026	0.185	0.023	0.381	0.027	0.027
(c) $\Omega/4$	H	0.427	0.022	0.210	0.021	0.364	0.027	0.027
	F	0.435	0.025	0.178	0.026	0.387	0.029	0.029

Tableau 5
Sensibilité de \hat{P}_j pour le choix de $\eta_j^{(0)}, \nu_j^{(0)}$ et $\eta_{4j}^{(0)}, \nu_{4j}^{(0)}$ pour les trois niveaux de l'IMC

groupe des personnes plus âgées (45+) pour les trois niveaux de l'IMC

Race	Sexe	\hat{P}_1	$\hat{P}_1 - \hat{P}_1^{(1)}$	\hat{P}_2	$\hat{P}_2 - \hat{P}_2^{(1)}$	\hat{P}_3	$\hat{P}_3 - \hat{P}_3^{(1)}$
------	------	-------------	-------------------------------	-------------	-------------------------------	-------------	-------------------------------

(a) 42	B	H	0.030	0.005	0.306	0.018	0.664	0.018
	F	0.081	0.002	0.436	0.004	0.483	0.004	0.004
(b) Ω	H	0.031	0.005	0.292	0.016	0.677	0.016	0.016
	F	0.063	0.002	0.443	0.006	0.494	0.005	0.005
(c) $\Omega/4$	H	0.073	0.015	0.359	0.011	0.568	0.019	0.019
	F	0.066	0.012	0.237	0.018	0.697	0.019	0.019

(a) 42	B	H	0.031	0.005	0.293	0.018	0.676	0.019
	F	0.073	0.015	0.359	0.011	0.568	0.019	0.019
(b) Ω	H	0.053	0.010	0.317	0.018	0.630	0.019	0.019
	F	0.065	0.013	0.221	0.022	0.714	0.025	0.025

Tableau 1 : $\Omega = (\eta^{(0)}, \nu^{(0)}, \eta_{41}^{(0)}, \nu_{41}^{(0)}, \eta_{42}^{(0)}, \nu_{42}^{(0)}, \eta_{43}^{(0)}, \nu_{43}^{(0)})$

Notas : 1 : Le modèle de non-réponse non-ignorable est appliqué au groupe des personnes plus âgées.

Notas : 2 : Le modèle de non-réponse non-ignorable est appliqué au groupe des personnes plus âgées.

Markov pour obtenir des estimations de la distribution a posteriori des paramètres. Notre méthode consiste à utiliser un échantillonneur de Metropolis-Hastings (MH) pour obtenir des échantillons à partir des équations en (8) et en (9) à partir desquels nous établissons des inférences a posteriori au sujet des valeurs de \mathbf{p}_i et δ_i .

3.2 Calculs

Pour le modèle de non-réponse ignorable, il convient de représenter la fonction de densité a posteriori comme étant

$$f(\mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 | \mathbf{y}, \mathbf{r}) = \prod_{i=1}^I \{ f_1(\mathbf{p}_i | \mathbf{y}, \mathbf{r}, \boldsymbol{\mu}_1, \boldsymbol{\tau}_1) f_2(\boldsymbol{\pi}_i | \mathbf{y}, \mathbf{r}, \boldsymbol{\mu}_2, \boldsymbol{\tau}_2) \}$$

où $f_1(\cdot)$ est la densité Dirichlet,

dont les distributions a priori sont $p(\boldsymbol{\mu}_3, \boldsymbol{\tau}_3)$ et $p(\boldsymbol{\mu}_4, \boldsymbol{\tau}_4)$. Par conséquent, nous obtenons f_1, \dots, f_{j+1} par le biais du noyau de Gibbs, tandis que nous obtenons f_{j+2} au moyen de l'algorithme MH (Nandram 1998). Nous obtenons les variables latentes \mathbf{z}_{ij} par le biais d'une des densités a posteriori subordonnées à l'algorithme MH. Un diagramme de Nous avons prélevé 5 500 itérations, rejete les 500 premières et conservé toutes les cinquinièmes (qui sont corrélées). Cette stratégie nous a permis d'éliminer l'auto-corrélation entre les itérations et d'obtenir de bonnes probabilités (0,25-0,50) pour les étapes Metropolis. Pour ce qui est du calcul, nous avons d'abord établi les hyperparamètres $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\nu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{\nu}_2^{(0)}, \boldsymbol{\mu}_3^{(0)}, \boldsymbol{\nu}_3^{(0)}, \boldsymbol{\mu}_4^{(0)}, \boldsymbol{\nu}_4^{(0)}$ pour $j = 1, \dots, J$ équivalant à 0. Puis, nous avons exécuté notre algorithme MH pour obtenir les échantillons a posteriori de $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3$ et $\boldsymbol{\tau}_4$, $j = 1, \dots, J$. Pour que les densités a posteriori soient appropriées, nous estimons les a priori gamma par rapport aux échantillons a posteriori pour $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3$ et $\boldsymbol{\tau}_4$, $j = 1, \dots, J$. Ces valeurs figurent au tableau 3. Enfin, à partir de ces a priori, nous avons exécuté notre algorithme de manière à obtenir les échantillons a posteriori. Plus précisément, nous avons obtenu les itérations $M = 1,000$ ($\mathbf{p}_i^{(h)}, \delta_i^{(h)}$), $h = 1, \dots, M$, $i = 1, \dots, c$. L'inférence au sujet des \mathbf{p}_i, δ_i et toute fonction les concernant peut être obtenue à partir de ces itérations d'une manière simple.

Tableau 3

Les estimations de $\eta^{(0)}$ et $\nu^{(0)}$ correspondant aux densités gamma sur $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3$ pour 45+ et $\boldsymbol{\tau}_3, \boldsymbol{\tau}_4, \boldsymbol{\tau}_{42}, \boldsymbol{\tau}_{43}$ pour 45- selon la race et le sexe

Age	45-		45+	
	$\boldsymbol{\tau}_3$	$\boldsymbol{\tau}_4$	$\boldsymbol{\tau}_3$	$\boldsymbol{\tau}_4$
Race	$\eta^{(0)}$	$\nu^{(0)}$	$\eta^{(0)}$	$\nu^{(0)}$
B	3,698	2,341	3,085	2,685
H	0,036	0,071	0,163	0,009
F	0,030	0,059	0,072	0,008
N	4,948	2,922	3,156	2,404
H	0,068	0,096	0,169	0,107
F	3,745	3,084	1,893	3,292
$\nu^{(0)}$	0,055	0,036	0,049	0,009
$\eta^{(0)}$	0,036	0,068	0,107	0,036
$\nu^{(0)}$	4,488	2,404	5,971	4,376

$$\left\{ \begin{aligned} & \left\{ n_z (\ell^I d(\ell^J \mathbf{x} - \mathbf{l}))_{\ell_K} (\ell^J d \ell^I \mathbf{x}) \right\}_{\ell}^{I=f} \\ & \left(\begin{array}{c} \ell^I z \dots \ell^I z \\ \ell^I \mathbf{d} - \ell^I \mathbf{u} \end{array} \right) \left(\begin{array}{c} \ell^I K \dots \ell^I K \\ \ell^I \mathbf{d} \end{array} \right) \left(\begin{array}{c} \ell^I \mathbf{d} \\ \ell^I \mathbf{u} \end{array} \right) \end{aligned} \right\}_{\mathcal{O}}^{I=1} = (\mathbf{x}' \mathbf{d} | \mathbf{z}' \mathbf{t}' \Delta) f$$

La fonction de vraisemblance en ce qui concerne le modèle de non-réponse ignorable est la suivante

$$\left\{ \left\{ {}_{1_A-1_U}({}^1\mathcal{U}-1) {}^1\mathcal{U} \begin{pmatrix} {}^1\mathcal{A} \\ {}^1\mathcal{U} \end{pmatrix} \right\} \prod_c^{1=1} = ({}^1\mathcal{U}, \mathbf{d} \mid \mathbf{1}, \mathbf{1}) f \right.$$

$$\cdot \left\{ i_{I - i_u + I_K}^{I_K} d \right\} \prod_{I=I}^{I=f} \left(\begin{matrix} I_K \dots I_K \\ I_A \end{matrix} \right) \left\{ \prod_{J=J}^{J=f} \right\}$$

Nous examinons l'inférence au sujet des $p_{ij}^{(t)}$, de la proportion des personnes au $j^{\text{ème}}$ niveau de l'IMC dans le $i^{\text{ème}}$ comté, de même que la probabilité de réponse,

$$g_i = \sum_{j=1}^J \pi_{ij} p_{ij}^{(t)}, i = 1, \dots, C.$$

$$\cdot \left\{ \left({^f v_2} \right)_{(0)}^{^f v_{1-}} \right\} d x \varrho \left. \right|_{1-}^{^f v_{11}} \left\{ \prod_{f=1}^{1= f} \left\{ \left(\varepsilon_2 \right)_{(0)}^{z_{1-}} \right\} d x \varrho \left. \right|_{1-}^{z_{11}} \right\} \times$$

De même, la fonction de vraisemblance augmentée (c'est-à-dire comprenant les \mathbf{z}_i) pour le modèle de non-réponse non-ignorable est la suivante

$$(8) \cdot \left\{ \left({}^{12} \mathbf{1}_{(0)} {}^{12} \mathbf{1} - \right) d \mathbf{x}_{\mathbf{1} - {}^{12} \mathbf{0}} \right\} \left\{ \left({}^{12} \mathbf{1}_{(0)} {}^{12} \mathbf{1} - \right) d \mathbf{x}_{\mathbf{1} - {}^{12} \mathbf{0}} \right\} \times$$

$$Y_{ijk} = \begin{cases} 1, & \text{si la personne } k \text{ appartient au niveau IMC } j \text{ dans le comté } i \text{ à répondu} \\ 0, & \text{si la personne } k \text{ appartient au niveau IMC } j \text{ dans le comté } i \text{ n'a pas répondu.} \end{cases}$$

Nous utilisons une structure probabilistique pour modéliser les valeurs de x_{ijk} et Y_{ijk} . Dans notre application, il y a $c = 34$ comtés et $J = 3$ niveaux d'IMC.

3.1 Modèles de non-réponse non-ignorable et ignorable

Pour ces deux modèles, nous avons

$$(1) \quad \mathbf{x}_k | \mathbf{p}_i \sim \text{Multinomial}(1, \mathbf{p}_i)$$

où p_{ij} constitue la probabilité qu'une personne dans le $i^{\text{ème}}$

comté appartienne au $j^{\text{ème}}$ niveau de l'IMC. Nous passons maintenant à la description des autres volets des modèles de non-réponse non-ignorable et ignorable.

D'abord, nous décrivons le modèle de la non-réponse ignorable. Soit π_i dénote la probabilité qu'une personne dans le $i^{\text{ème}}$ comté réponde (c'est-à-dire que la probabilité de réponse dépend essentiellement du comté). Alors, nous présumons que

$$(2) \quad Y_{ijk} | \pi_i \sim \text{Bernoulli}(\pi_i).$$

À la deuxième étape, soit $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iJ})'$, nous prenons

$$(3) \quad \mathbf{p}_i | \mu_i, \tau_i \sim \text{Dirichlet}(\mu_i, \tau_i),$$

$$(4) \quad \pi_i | \mu_{21}, \tau_{21} \sim \text{Beta}(\mu_{21}, \tau_{21}, (1 - \mu_{21}), \tau_{21})$$

$$p(\mathbf{p}_i | \mu_i, \tau_i) \prod_{j=1}^J p_{ij}^{\mu_{ij} \tau_{ij} - 1} / D(\mu_i, \tau_i, 0 < p_{ij} < 1, \sum_j p_{ij} = 1$$

et

$$D(\mu_i, \tau_i) = \prod_{j=1}^J \Gamma(\mu_{ij} \tau_{ij}) / \Gamma(\tau_i), 0 < \mu_{ij} < 1, \sum_j \mu_{ij} = 1.$$

Les composantes de μ_i sont les principales moyennes a

priori des composantes correspondantes de \mathbf{p}_i , et la valeur de τ_i peut être interprétée comme une taille d'échantillon a priori. On peut faire des interprétations semblables pour μ_{21} et τ_{21} pour π_i . Par conséquent, l'hypothèse (3) exprime des similarités entre les proportions de cellules \mathbf{p}_i et l'hypothèse (4) exprime des similarités entre les probabilités de réponse π_i . C'est cette structure qui dicte l'emprunt d'information entre les comtés c .

Puis, nous décrivons le modèle de non-réponse non-ignorable. Soit π_{ij} dénote la probabilité qu'une personne dans le $i^{\text{ème}}$ comté réponde au $j^{\text{ème}}$ niveau de l'IMC (c'est-à-dire que la probabilité de réponse dépend non

seulement du comté mais aussi du niveau de l'IMC). Alors, nous présumons que

$$(5) \quad Y_{ijk} | \{x_{i1k}, \dots, x_{iJk}\}, \pi_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

où $x_{ijk} = 1, x_{ijk} = 0, j \neq i$ pour $j, i' = 1, 2, \dots, J$. Soit $\mu_j = (\mu_{j1}, \mu_{j2}, \dots, \mu_{jJ})'$, à la deuxième étape nous prenons également

$$(6) \quad \mathbf{p}_i | \mu_j, \tau_j \sim \text{Dirichlet}(\mu_j, \tau_j)$$

et

$$(7) \quad \pi_{ij} | \mu_{4j}, \tau_{4j} \sim \text{Beta}(\mu_{4j}, \tau_{4j}, (1 - \mu_{4j}), \tau_{4j}), j = 1, \dots, J.$$

Comme pour les hypothèses en (3) et en (4), les hypothèses en (6) et en (7) expriment des similarités entre les comtés. Nous observons que les paramètres de réponse π_{ij} sont faiblement identifiables (c'est-à-dire que les estimations ne sont pas fiables). Toutefois, le modèle de sélection joue en notre faveur parce que la densité conjointe de x_{ijk} et $Y_{ijk} = (Y_{i1k}, \dots, Y_{iJk})'$ relie les valeurs de p_{ij} et π_{ij} . De fait, il s'agit là d'un avantage par rapport à la méthode des modèles combinés.

Pour que l'analyse bayésienne soit complète, nous prenons à la troisième étape les densités a priori qui s'appliquent aux hyper-paramètres comme suit. Pour le modèle de non-réponse ignorable, les densités a priori sont les suivantes

$$\mu_i \sim \text{Dirichlet}(1, 1, \dots, 1), \mu_{21} \sim \text{Beta}(1, 1),$$

$$\tau_i \sim \text{Gamma}(\eta_i^{(0)}, \nu_i^{(0)}) \text{ et } \tau_{21} \sim \text{Gamma}(\eta_{21}^{(0)}, \nu_{21}^{(0)}),$$

où (soit t dénote τ_i ou τ_{21} , soit η_i ou η_{21} , et b soit $\nu_i^{(0)}$ ou $\nu_{21}^{(0)}$) t signifie que $f(t) = b a^t a^{-b} / \Gamma(a)$, $t > 0$ et $f(t) = 0$ autrement. Les hyper-paramètres $\eta_i^{(0)}, \nu_i^{(0)}, \eta_{21}^{(0)}$ et $\nu_{21}^{(0)}$ doivent être définis. La partie correspondante du modèle de non-réponse non-ignorable est la suivante

$$\mu_j \sim \text{Dirichlet}(1, 1, \dots, 1), \mu_{4j} \sim \text{Beta}(1, 1),$$

$$\tau_j \sim \text{Gamma}(\eta_j^{(0)}, \nu_j^{(0)}) \text{ et } \tau_{4j} \sim \text{Gamma}(\eta_{4j}^{(0)}, \nu_{4j}^{(0)}), j = 1, \dots, J.$$

De nouveau, les hyper-paramètres $\eta_j^{(0)}, \nu_j^{(0)}, \eta_{4j}^{(0)}$ et $\nu_{4j}^{(0)}$, $j = 1, \dots, J$, doivent être définis. On peut utiliser d'autres densités a priori, comme des a priori de réduction, mais vraisemblablement celles-ci donneraient une inférence similaire à ce que notre analyse de la sensibilité a montré à la section 4.

Le modèle hiérarchique a une propriété bien attrayante qui consiste à établir des corrélations entre les variables. Par exemple, dans notre application, les équations (1), (2), (3) et (4) rendent la valeur de (x_{ij}, y_{ij}) en équicorrélation d'une

n'est pas une procédure aléatoire dans une population composée à la fois d'hommes et de femmes.

Tableau 1

Nombre de personnes dans chaque niveau de l'IMC et nombre de non-répondants (Non) selon l'âge, la race et le sexe dans les 34 régions

Âge	Race	Sexe	IMC		
			1	2	3
45-	B	H	1 098	651	597
	B	F	845	434	380
	N	H	1 198	713	665
45+	B	H	46	439	1 014
	B	F	51	223	365
	N	H	79	470	942
		F	48	169	552
		Non	6		

Nota : IMC (1=moins de 20; 2 = entre 20 et 25; 3 = plus de 25)

Âge (plus jeune que 45 ans = 45-; 45 ans ou plus = 45+) Race (Blanc = B; tous les autres = N) Sexe (Homme = H; Femme = F)

Un des aspects importants de nos travaux concerne l'esti-

mation régionale. Parce que nous tenons compte de l'inté-
rence pour chaque domaine d'âge, de race et de sexe séparé-
ment d'une région géographique à l'autre (comtés), les
échantillons de certaines de ces régions peuvent être très
petits. Par conséquent, les techniques d'estimation régionale
doivent permettre d'estimer les paramètres correspondant à
ces petites régions. En particulier, nous devons « emprunter
de l'information » des plus grandes régions pour accroître la
fiabilité des estimations des petites régions. Au tableau 2, il
y a huit exemples qui illustrent la nécessité d'adopter des
techniques adaptées aux petites régions. Nous avons
sélectionné huit comtés qui ont des petites domaines; toutes
les cellules comptent au plus six unités et bon nombre
d'entre elles en comptent aussi peu qu'une (1) une d'elles est
0 pour le groupe des 45+). Nous présenterons des esti-
mations globales, de même que les estimations pour les
quatre premiers exemples (45-). Souliignons que, en compa-
raison avec les comtés complexes des cellules, le nombre de non-
répondants est important pour deux d'entre elles (14 et 10
non-répondants).

Précisons que l'objectif n'est pas d'effectuer une analyse
exhaustive des données de la NHANES III même s'il s'agit
la d'une analyse approximative de ces données. Notre
méthode est suffisamment générale pour nous permettre
d'analyser la non-réponse multinomiale de nombreuses
régions, dont certaines sont petites. C'est pour ces petites
régions que nous avons conçu cette méthode de modélisa-
tion. Par conséquent, dans le présent article, nous utilisons
les données de la NHANES III pour illustrer notre méthode.
Selon notre méthode, nous tenons compte de chaque
domaine séparément et procédons à un « emprunt
d'information » auprès des 34 régions (comtés) pour analyser
les données de l'IMC. Par conséquent, il y a huit analyses
distinctes portant sur 34 régions, dont certaines sont petites.

Tableau 2
Nombre de personnes dans chaque niveau de l'IMC et nombre de non-répondants (Non) dans les huit exemples (Ex) de petits domaines âge-race-sexe de différents comtés

Ex	Âge	Race	Sexe	IMC		
				1	2	3
1	45-	B	H	1	3	1
		B	F	3	4	1
		N	H	5	6	10
2		B	F	3	1	1
		N	H	5	5	6
		N	F	3	4	10
3	45+	B	H	1	2	6
		B	F	1	3	4
		N	H	3	3	5
4		B	F	2	0	1
		N	H	3	3	5
		N	F	3	3	5
5	45-	B	H	2	0	1
		B	F	1	3	4
		N	H	3	3	5
6		B	F	1	3	4
		N	H	3	3	5
		N	F	3	3	5
7	45+	B	H	2	0	1
		B	F	1	3	4
		N	H	3	3	5
8		B	F	2	0	1
		N	H	3	3	5
		N	F	3	3	5

Nota : IMC (1=moins de 20; 2 = entre 20 et 25; 3 = plus de 25)

Âge (plus jeune que 45 ans = 45-; 45 ans ou plus = 45+) Race (Blanc = B; tous les autres = N) Sexe (Homme = H; Femme = F)

3. MÉTHODOLOGIE QUI S'APPLIQUE AU MODÈLE HIÉRARCHIQUE MULTINOMIAL

Nous proposons un modèle pour chacun des huit domaines âge-race-sexe, mais qui s'applique à tous les comtés simultanément. Cependant, les modèles s'inscrivent dans deux grandes classes. Nous utiliserons un modèle de non-réponse non-ignorable pour le groupe des plus jeunes et un modèle de non-réponse ignorable pour le groupe des personnes plus âgées, vu que le taux de non-réponse pour ce dernier groupe est négligeable. Bien entendu, cela vaut la peine de comparer le modèle de non-réponse ignorable et le modèle de non-réponse non-ignorable pour le groupe des personnes plus jeunes. Nous montrons comment combiner les groupes plus tard au moyen de la régression logistique, bien qu'il ne s'agisse pas ici de le but principal du présent article.

Pour chaque groupe âge-race-sexe, la $k^{\text{ième}}$ personne dans le $j^{\text{ième}}$ comté appartient à l'un des niveaux de l'IMC $j^{\text{ième}}$. Alors pour la $k^{\text{ième}}$ personne dans le $j^{\text{ième}}$ comté, la variable de la caractéristique au $j^{\text{ième}}$ niveau de l'IMC est définie comme suit,

$$x_{ijk} = (x_{1jk}, \dots, x_{Ijk}, \dots, x_{I'jk}, \dots, x_{I''jk})', i = 1, \dots, c, k = 1, \dots, n_j$$

où chaque $x_{ijk} = 0$ ou 1, $j = 1, \dots, J$, et $\sum_{j=1}^J x_{ijk} = 1$. La variable de la réponse, y_{ijk} , est définie pour chaque domaine âge-race-sexe

Une des variables intéressantes dans la NHANES III est l'IMC, un indice du poids rajustés en fonction de la taille (Kg/m^2), qui classe dans des catégories larges l'obésité au sein des groupes d'âge, de race et de sexe (Kuczmarski, Carroll, Flegal et Troiano 1997) comme faible pourcentage d'adiposité (niveau 1 : $\text{IMC} < 20$), pourcentage d'adiposité santé (niveau 2 : $20 \leq \text{IMC} < 25$), pourcentage d'adiposité non santé (niveau 3 : $\text{IMC} \geq 25$). Nous utilisons cette vaste classification pour chacun des huit groupes d'âge, de race et de sexe.

Pourtant que de se limiter à une analyse des données nominales, on peut faire une analyse où l'IMC est considérée comme une variable continue. Comme certains renseignements sont perdus dans le cadre de la discrétisation des valeurs de l'IMC, une analyse qui s'appuie sur des modèles continus de l'IMC donnerait également des données approximatives; c'est pourquoi il faut chercher à obtenir une transformation appropriée. Dans l'analyse finale, les individus n'ont qu'à savoir quelle proportion du public s'inscrit dans chaque niveau de l'IMC pour pouvoir dire à leurs patients où ils se situent par rapport à l'obésité.

L'analyse des données de l'IMC selon des méthodes de données nominales n'est pas rare. Par exemple, Malec, Davis et Cao (1999) ont décrit une analyse de Bayes fondée sur une méthode empirique de Bayes des données de la NHANES III. Ils classaient les personnes de plus de 20 ans comme normales si leur IMC était inférieur à un certain seuil établi selon le sexe. Il s'agit là d'une application d'un analyse bayésienne des données binaires. Leur classification est toutefois quelque peu restreinte (voir Kuczmarski et coll., 1997). En tenant compte des données multinomiales, nous avons généralisé l'analyse de Malec et coll. (1999). En fait, ils n'ont pas fourni de modèle de non-réponse non-ignorable.

Contrairement à Schaffer et coll. (1996), nous incluons la mise en grappes au niveau du comté, même s'il faudrait le faire au niveau du ménage. Pour les données complètes, y a 6 440 ménages. Dans 52,1 % de ces ménages, une personne faisait partie de l'échantillon, dans 22,5 %, deux personnes et dans 21,4 %, au moins trois personnes. Nous avons calculé le coefficient de corrélation pour les valeurs de l'IMC en fonction du couplage des membres au sein des ménages (voir Rao 1973, page 199). C'est le chiffre 0,19 qui indique qu'à titre de première approximation on peut ne pas tenir compte du regroupement en grappes à l'intérieur des ménages.

Au tableau 1 figure le nombre de répondants à chaque niveau de l'IMC pour chaque domaine âge-race-sexe et 34 comtés (population d'au moins 500 000 habitants). Le modèle des répondants diffère considérablement selon l'âge. Le taux de non-réponse dans le cas du groupe de personnes plus âgées (+45) est négligeable. Par conséquent, en ce qui a trait à la non-réponse, il faut surtout s'attarder au groupe de personnes plus jeunes (-45). De plus, le taux de réponses est plus élevé chez les femmes que chez les hommes. Nous observons que la procédure de sélection

2. LES DONNÉES DE LA NHANES III ET LA

NON-RÉPONSE

nous décrivons la NHANES III. Dans la section 3, nous abordons le modèle bayésien appliqué à la non-réponse non-ignorable. Nous appliquons plus précisément un modèle bayésien hiérarchique en trois étapes aux données multinomiales de la NHANES III pour étudier le problème de la non-réponse. Dans la section 4, nous décrivons une analyse des données de la NHANES III dans le cadre de laquelle nous incluons une analyse de régression visant à combiner tous les domaines liés à l'âge, à la race et au sexe. Dans la section 5, nous décrivons une étude de simulation qui nous permet d'évaluer le rendement de notre modèle. Enfin, nous concluons à la section 6.

La NHANES III est une des enquêtes périodiques servant à évaluer un aspect de la santé de la population américaine (National Center for Health Statistics 1994). Ce qui intéresse comme champ d'études c'est la non-réponse à l'indice de masse corporelle (IMC) dans la NHANES III. Les données dont nous sommes servis proviennent de cette enquête et ont été recueillies entre octobre 1988 et septembre 1994. À la section 2.1, nous décrivons les données réelles, et à la section 2.2, nous décrivons les données que nous analysons.

2.1. Les données de la NHANES III

La NHANES III comporte deux parties. La première partie est l'interview des personnes échantillonnées en vue d'obtenir des renseignements personnels et la deuxième partie est l'examen des personnes échantillonnées. Une ou plusieurs personnes des ménages échantillonnés ont été placées dans un certain nombre de sous-groupes selon leur âge, leur race et leur sexe. Certains sous-groupes ont été échantillonnés à différents taux. Les personnes échantillonnées devaient se rendre à un centre d'examen mobile (MEC) pour un examen physique. Ceux qui ne s'y rendaient pas recevaient une visite de l'examinateur. Des précisions sur le plan de sondage de la NHANES III sont disponibles (National Center for Health Statistics 1992). Dans notre modèle, nous incorporons les caractéristiques du plan associées à la mise en grappes.

Les raisons principales de non-réponse dans le cas de la NHANES III sont les suivantes : non intéressé, n'a pas le temps/en conflit avec le travail, préoccupation/métier, ne me dérangez pas et des raisons de santé. Le taux de non-réponse des jeunes personnes est très élevé parce que les parents, en particulier les mères plus âgées d'un enfant unique, étaient très protecteurs de leurs bébés, et refusaient que ceux-ci ne quittent la maison pour le MEC. Les travailleurs sur le terrain ont souvent observé que les personnes obèses avaient tendance à essayer de se soustraire à l'examen médical. La non-réponse pouvait ainsi ne pas être aléatoire et, du coup, mériter qu'on s'y attarde.

modèle de non-réponse subordonné aux données hypothétiques. On a conçu cette approche dans le but d'étudier les problèmes en matière de prélèvement d'échantillon (voir Heckman 1976 et Olson 1980). Dans l'approche des modèles combinés, les répondants et les non-répondants sont modélisés séparément, et la réponse finale est obtenue au moyen d'un mélange probabilistique des deux modèles. Stasny (1991) a utilisé un modèle empirique de Bayes pour étudier la victimisation dans le cadre de la National Crime Survey et appliqué la méthode de sélection. Dans le cadre de cette analyse, elle a regroupé les données binomiales de plusieurs domaines, dont certains ont peu de cellules. Il s'agit pour l'essentiel d'un exercice d'estimations régionales. Albert et Gupta (1985) ont présenté une méthode connue fondée sur une approximation pour obtenir une approche bayésienne pour une population comptant un seul domaine (voir également Kaufman et King 1973). Autrement dit, contrairement à Stasny (1991), ces auteurs n'ont pas effectué d'estimations régionales, et leur analyse à l'égard d'un seul domaine ne s'appuie sur aucune donnée d'autres domaines.

Vu que la méthode bayésienne permet d'intégrer d'autres renseignements au sujet des non-répondants, elle convient tout à fait à l'analyse de la non-réponse non-ignorable (Little et Rubin 1987, et Rubin 1987). Cependant, le principal problème est la façon de décrire la relation entre répondants et non-répondants. Au moyen de la méthode de sélection dans le cadre de la méthode empirique de Bayes (voir Deely et Lindley 1981), Stasny (1991) a d'abord estimé les hyperparamètres selon des méthodes du maximum de vraisemblance, puis a présupposé que ceux-ci étaient connus, ce qui a permis d'éliminer une certaine quantité de variabilité. Nous avons élargi le champ d'application de cette méthode dans deux directions.

Nous commençons par étudier les données multinomiales obtenues de façon indépendante auprès de plusieurs régions géographiques. Précisons que Basu et Pereira (1982) ont envisagé les données multinomiales manquantes dues à la non-réponse se rapportant à un seul domaine selon un modèle multinomial de Dirichlet alors que l'on présuppose que les hyper-paramètres sont connus. Récemment, Forster et Smith (1998) ont utilisé des modèles log linéaires de Dirichlet de données multinomiales graphiques pour analyser les données de l'enquête par panel à l'élection générale britannique. La encore, nous présupposons que les hyper-paramètres sont connus et nous utilisons un modèle comportant un seul domaine. Puis, nous employons une approche bayésienne complétée pour tenir compte de la non-réponse multinomiale de plusieurs régions. Nous n'estimons pas les hyper-paramètres en fonction des données.

En résumé, nous élaborons un modèle de non-réponse non-ignorable qui permet de regrouper les données qui concernent de nombreuses petites régions et nous notions que nous pouvons en servir dans d'autres applications. Le reste de l'article est structuré comme suit. Dans la section 2,

Un modèle bayésien hiérarchique de non-réponse pour les données multinomiales des petites régions

BALGOBIN NANDRAM, GEUNSHIK HAN et JAI WON CHOI

RÉSUMÉ

L'analyse des données d'enquête de différentes régions géographiques, qui sont des données polychotomiques, se fait facilement au moyen de modèles bayésiens hiérarchiques même si certaines de ces régions présentent des cellules avec petits chiffres. Il y a toutefois des problèmes quand les données d'enquête sont incomplètes en raison de non-réponses, en particulier quand les caractéristiques des répondants diffèrent de celles des non-répondants. En présence de non-réponses, nous appliquons la méthode de sélection pour l'estimation parce qu'elle nous permet de procéder à des inférences à l'égard de tous les paramètres. Plus précisément, nous décrivons un modèle bayésien hiérarchique pour l'analyse de la non-réponse multinomiale non-ignorable de diverses régions géographiques, dont certaines sont petites. Pour le modèle, nous utilisons une densité à priori Dirichlet pour les probabilités multinomiales et une densité à priori bêta pour les probabilités de fiabilité des estimations des paramètres du modèle qui s'appliquent aux petites régions. Parce que la densité conjointe a position de tous les paramètres est complexe, l'inférence se fonde sur l'échantillonnage d'après la méthode de Monte Carlo (MC) tirées de la troisième édition de la National Health and Nutrition Examination Survey (NHANES III). À des fins de clarté, l'IMC est classé selon trois niveaux naturels pour chacun des huit domaines âge-race-sexe et 34 comtés. Nous évaluons le rendement de notre modèle à partir des données de la NHANES III et d'exemples simulés qui nous indiquent que notre modèle fonctionne passablement bien.

MOTS CLÉS : Variable latente; échantillonnage de Metropolis-Hastings; non-réponse non-ignorable; méthode de sélection; petite région.

1. INTRODUCTION

Les taux de non-réponse dans de nombreuses enquêtes ont augmenté régulièrement (De Heer 1999; et Groves et Couper 1998), ce qui accentue le problème de la non-réponse. Les réponses sont polychotomiques dans de nombreuses enquêtes. Par exemple, dans la troisième édition de la National Health and Nutrition Examination Survey (NHANES III), nous pouvons estimer les proportions des personnes appartenant aux trois niveaux de l'indice de masse corporelle (IMC), bien que l'IMC soit une variable continue. L'objectif du présent article est de décrire un nouveau modèle bayésien hiérarchique qui nous permet d'étudier la non-réponse multinomiale non-ignorable pour les petites régions, et de l'appliquer aux données de l'IMC de la NHANES III.

Rubin (1987), ainsi que Little et Rubin (1987) décrivent deux types de modèles qui diffèrent selon que l'on doive tenir compte ou non de la réponse. Dans le modèle de la non-réponse ignorable, la répartition de la variable qui nous intéresse à l'égard des répondants est la même que la répartition de la variable dans le cas des non-répondants; les paramètres des répartitions de la variable et de la réponse doivent être distincts (voir Rubin 1976). Tous les autres modèles de non-réponse doivent être tenus en compte. Nous

utilisons à la fois des modèles de non-réponse non-ignorable et ignorable pour nos données parce que pour certains domaines il n'y a aucun non-répondant. Crawford, Johnson et Laird (1993) ont analysé les données de la Harvard Medical Practice Survey à partir de modèles de non-réponse non-ignorable. Stasny, Kadam et Fritsch (1998) ont établi, à partir d'un modèle bayésien hiérarchique, les probabilités de déclarer un prévenu coupable ou non lors d'un procès particulier où les points de vue des non-répondants différaient de ceux des répondants à l'égard de la peine de mort. Park et Brown (1994) ont appliqué une pseudo méthode bayésienne (Baker et Laird 1988), tandis que Park (1998) a appliqué une méthode selon laquelle des observations a priori sont attribuées à la fois aux cellules observées et aux cellules non observées dans le but d'estimer les cellules manquantes d'un tableau de variables nominales à plusieurs dimensions dans le cadre de la non-réponse non-ignorable. Notre approche diffère de celle de ces auteurs. Nous décrivons l'estimation régionale à l'égard des données multinomiales et nous appliquons les techniques au moyen des méthodes de Monte Carlo à chaînes de Markov. Ainsi, nous pouvons inclure à nos modèles toutes les sources de variabilité. On peut modéliser la non-réponse de deux façons. L'approche de la sélection est utilisée dans le cas de données complètes hypothétiques à laquelle est ajouté un

calculé \sqrt{R} (Gelman et Rubin, 1992) pour \bar{Q} (s, t) dans (3), pour des séries de 1 000, 2 000 et 4 000 itérations, après respectivement. Après 2 000 itérations, avec 2 000 itérations de fiabilisation, nous avons observé que $\sqrt{R} \leq 1,010$ dans tous les cas étudiés, y compris ceux du tableau 3. Nous pensons que ce niveau d'exactitude est acceptable pour l'approximation des modes et des variances.

BIBLIOGRAPHIE

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley-Interscience.

BESAG, J. (1974). Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society, B*, 36, 2.

BISHOP, Y. M. M., FIENBERG, S. E. et HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.

CHEN, J. et SHAO, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 365-369.

CHEN, J., et SHAO, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 2.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3^e Edition. Wiley.

DEMMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-22.

FAY, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 227-232.

FAY, R. E. (1999). Theory and application of nearest neighbor imputation in census 2000. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 112-121.

FAY, R. E., et TOWN, M. K. (1998). Variance estimation for the 1998 census dress rehearsal. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 605-610.

GELMAN, A., et RUBIN, D. B. (1991). Simulating the Posterior Distribution of Loglinear Contingency Table Models. Rapport technique non publié, Harvard University.

GELMAN, A., et RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 4.

KOSTANICH, D. L. (1999). DSSD Census 2000 Dress Rehearsal Memorandum Series #A, US Bureau of the Census.

KOVAR, J. G., et WHITTRIDGE, P. J. (1995). Imputation of Business Survey Data. *Business Survey Methods*, (Eds. Cox, D., Binder, Chinnappa, Christianson, M., Colledge et Kott). Wiley.

LARSEN, M. D. (1996). *Bayesian Approaches to Finite Mixture Models*. Dissertation de doctorat, Department of Statistics, Harvard University.

LITTLE, R. J. A., et RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

MENG, X. M. (1994). Multiple-imputation inferences with uncongential sources of input. *Statistical Science*, 9, 4.

RAO, J. N. K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4.

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 434.

SANDE, I. G. (1981). Imputation in surveys: coping with reality. *Techniques d'enquête*, 7, 21-43.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

THIBAUDEAU, Y. (1988). *Approximating the Moments of a Multimodal Posterior Distribution with the Method of Laplace*. Dissertation de doctorat, Department of Statistics, Carnegie Mellon University.

TREAT, J. B. (1994). *Summary of the 1990 Census Imputation Procedures for the 100 % Population and Housing Items*. DSSD REX Memorandum Series BB-11, US Bureau of the Census.

WILSON, E. B. (1998). Communication to Dan E. Philip. Housing and Household Economics Statistics Division, US Bureau of the Census.

ZANUTTO, E., et ZASLAVSKY, A. M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 608-613.

ZANUTTO, E., et ZASLAVSKY, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse follow-up. *Proceedings of the Annual Research Conference*, U.S. Department of Commerce, Bureau of the Census, 673-686.

Tableau 3

EMV, mode a posteriori (approximatif) et écart-type pour les probabilités conditionnelles d'êtres d'origine hispanique, étant donné la race, pour quatre choix de distribution a priori

Race	EMV	Mode	E.-T.	Mode	E.-T.	Mode	E.-T.	Mode	E.-T.
Blanc	0,1784	0,178	0,01195	0,184	0,01247	0,18	0,01219	0,188	0,01186
Noir	0,07428	0,069	0,02272	0,081	0,0233	0,12	0,02428	0,16	0,02782
Asiatique	0,09113	0,105	0,04086	0,108	0,0455	0,195	0,04881	0,276	0,04952
Autre	0,9662	0,966	0,01171	0,964	0,01347	0,95	0,01495	0,93	0,01666

Tableau 4
Chiffres de population et mesures de l'incertitude pour Sacramento

	Chiffre par imp. HDSH	Chiffre par imp. avec le modèle	EMV du chiffre attendu	Erreur du résiduel du modèle	Erreur de préd. de l'EMV	Erreur totale
Total	138271	138 271	138 271	0	0	0
Blancs	89 032	88 914	88 927,7	31,5	35,2	47,2
Noirs	19 962	19 943	19 952,9	14,9	16,5	22,3
Asiatiques	17 405	17 421	17 426,2	14	14,9	20,5
Autres	11 872	11 993	11 964,1	29,8	33,5	44,8
Hispaniques	21 024	21 050	21 038,1	10,3	10,6	14,7
Non-hispaniques	117 247	117 221	117 232,8	10,3	10,6	14,7
Propriétaires	70 054	70 022	70 026,3	42,8	43,3	60,9
Locataires	68 217	68 249	68 244,7	42,8	43,3	60,9
Blancs hispaniques	9 068	8 972	8 991,1	29,9	33,6	45
Blancs non hispaniques	79 964	79 942	79 936,6	15,4	15,7	22
Noirs hispaniques	605	612	608,6	11	12,6	16,7
Noirs non hispaniques	19 357	19 331	19 344,3	10,8	10,7	15,2
Asiatiques hispaniques	518	515	516,5	10	11,5	15,2
Asiatiques non hispaniques	16 887	16 906	16 909,7	10,4	10,3	14,6
Autres hispaniques	10 833	10 951	10 921,9	29,7	33,3	44,6
Autres non hispaniques	1 039	1 042	1 042,3	3,5	3,4	4,9
Blancs propriétaires	47 722	47 767	47 770,5	37,8	41,3	56
Blancs locataires	41 310	41 147	41 157,3	39	41,4	56,9
Noirs propriétaires	7 661	7 538	7 542,3	19,6	20,7	28,5
Noirs locataires	12 301	12 405	12 410,6	21,1	22,5	30,8
Asiatiques propriétaires	9 810	9853	9872,8	18,4	18,6	26,1
Asiatiques locataires	7595	7 568	7 553,4	18,2	18,8	26,1
Autres propriétaires	4 861	4 864	4 840,7	24,4	28,2	37,3
Autres locataires	7 011	7 129	7 123,4	25,4	28,6	38,2
Hispaniques propriétaires	9 409	9 434	9 402,2	19,5	20,9	28,6
Hispaniques locataires	11 615	11 616	11 629,9	20,1	21,4	29,4
Non hispaniques propriétaires	60 645	60 588	60 618	38,9	39,4	55,4
Non hispaniques locataires	56 602	56 633	56 614,8	38,7	39,6	55,4

qui ont déclaré être dans l'état s et $\Delta(s, \lambda)$ est biminimal $(\delta(\lambda), p_\lambda(s))$. En outre, posons que $S(s) = S^{obs}(s) + \sum_{\lambda=1}^L \delta(\lambda) p_\lambda(s)$, où $\hat{p}_\lambda(s)$ est l'EMV de $p_\lambda(s)$. Si nous traitons les λ comme des prédicteurs indépendants, comme dans le cas d'une régression, et puis que \hat{P} est asymptotiquement normal et que sa moyenne est \hat{P} , nous obtenons l'approximation pour grand échantillon qui suit pour l'EMV de $S(s)$ en vue de prédire $S(s)$.

$$E \left[\sum_{\lambda=1}^L \delta(\lambda) p_\lambda(s) - \Delta(s, \lambda) \right]^2 \middle| P$$

$$\approx V \left(\sum_{\lambda=1}^L \delta(\lambda) p_\lambda(s) \middle| P \right) + V \left(\sum_{\lambda=1}^L \Delta(s, \lambda) \middle| P \right). \quad (4)$$

Soit V_P et V_s , les première et deuxième variances dans

le membre droit de (4). Gelman et Rubin (1991), Larsen (1996) et Schaffer (1997, page 324) introduisent l'ajustement proportionnel itératif bayésien avec augmentation de données (APIBAD) pour simuler les distributions des estimateurs conformes au modèle pour V_P et V_s grâce à des simulations de la distribution a posteriori de $\sum_{\lambda=1}^L \delta(\lambda) p_\lambda(s)$ et de la distribution prédictive de $S(s)$, respectivement. En outre, nous obtenons une approximation par imputation au moyen de $A(f)$ en ajoutant un autre V_s^e à V_P dans (4) pour tenir compte du bruit supplémentaire du « jete de dés » compris dans $A(f)$.

6. MODÉLISATION ET ANALYSE DE SENSIBILITÉ

6.1 Un modèle d'indépendance conditionnelle pour Sacramento

En nous servant de la notation décrite à la section 3, les variables relatives aux chefs de ménage contenues dans Ψ sont la race, l'origine, le mode d'occupation du logement et le sexe. Pour la race, les catégories sont blanche, noire, hispanique et autre. Pour l'origine, elles sont hispanique et non hispanique. Pour le mode d'occupation du logement elles sont propriétaire et locataire. Pour le sexe, elles sont homme et femme. Les variables relatives au voisin dans Ξ sont la race, l'origine et le mode d'occupation du logement. Pour la race du voisin, les catégories sont noire et non noire, pour l'origine et le mode d'occupation du logement, les catégories sont les mêmes que pour le chef de ménage. Nous construisons Θ dans (3) en sélectionnant des interactions entre les variables comprises dans Ψ et Ξ . Pour nous assurer de l'équivalence entre (2) et (3), nous

6.2 Analyse de sensibilité et évaluation

À la section 7, nous utilisons l'erreur-type de la distribution prédictive de $S(s)$ pour approximer $\sqrt{V_P^e + V_s}$, c'est-à-dire l'erreur de $S(s)$ lors de la prédiction de $S(s)$ telle qu'établie dans (4), et nous supposons que $S(s) - S(s)$ est asymptotiquement normale. L'exactitude de cette approximation dépend de l'exactitude de l'approximation de la distribution de l'EMV \hat{P} au moyen de la distribution a posteriori de \hat{P} . Cette dernière approximation est asymptotiquement exacte lorsque le modèle est véritable, mais nous devons néanmoins déterminer la mesure dans laquelle ce résultat asymptotique est applicable lorsque l'échantillon est fini. Pour cela, nous examinons la sensibilité de la distribution a posteriori de \hat{P} en cas de changements a priori. Une faible sensibilité sous-entend que la

réponse à la question correspondante comprise dans Ψ . Nous complétons le modèle en sélectionnant des associations assurant la cohérence. Nous incluons les six interactions représentant les associations ayant trait à une paire de questions dans Ψ . Le tableau de contingence résultant contient 256 cellules et le modèle log-linéaire comprend 30 paramètres libres.

Ce modèle mène à une structure de transition à indépendance conditionnelle. Par exemple, subordonnée à la race du voisin le plus proche à réponse complète, la race du chef de ménage est indépendante du mode d'occupation du logement du voisin le plus proche à réponse complète. L'indépendance conditionnelle nous permet de combiner les renseignements sur les voisins provenant de voisins multiples pour produire un voisin le plus proche à réponse complète synthétique. Cette méthode assure que nous puissions utiliser toute l'information disponible provenant du voisin le plus proche, même si il n'a pas fourni une réponse à toutes les questions. Grâce à cette méthode, la structure de corrélation entre les questions auxqueltes a répondu le chef de ménage est maintenue lorsque l'on impute que la réponse à une seule question par chef de ménage. À Sacramento, parmi les 138 271 chefs de ménage, environ 0,1 % n'ont pas déclaré leur sexe, 3,5 % n'ont pas déclaré leur race, 2,9 % n'ont pas déclaré leur origine et 7,6 % n'ont pas déclaré leur mode d'occupation du logement. En outre, les réponses concernant la race et l'origine manquent toutes deux pour 0,49 % des chefs de ménage, celles pour la race et le mode d'occupation du logement manquent pour 0,48 %, et celles pour l'origine et le mode d'occupation du logement pour 0,69 %. Étant donné les faibles taux de réponses manquant conjointement, nous nous attendons à ce que le modèle donne de bons résultats.

Pour imputer les valeurs correspondant aux questions comprises dans la différence d'ensembles $\Psi - Z$ conformément à $A(f)$, nous jetons des dés pondérés par les valeurs de l'EMV de $P(s|t, Z, v)$ pour chaque chef de ménage dans l'état marginal v et dont le voisin le plus proche à réponse complète est dans l'état t . En vertu de nos présupposés, l'EMV de $P(s|t, Z, v)$ contient toute l'information disponible dans le fichier f pour les questions sans réponse. À la section suivante, nous formulons une fonction de vraisemblance pour $P(s|t, Z, v)$.

4. UNE FONCTION DE VRAISEMBLANCE POUR LES PROBABILITÉS DE TRANSITION

Soit $N(t, Z, v)$, le nombre de chefs de ménage qui ont répondu uniquement aux questions définissant l'état marginal v ne comportant que les questions comprises dans $Z \subset \Psi$, et dont le voisin le plus proche à réponse complète est dans l'état t . Soit N_v , un vecteur ayant pour composante les $N(t, Z, v)$, au niveau d'un secteur de recensement. Représentons par $P = [P(s|t)]$ le vecteur comprenant les $P(s|t)$ classés lexico-graphiquement selon t et s . Étant donné les présupposés décrits plus haut, nous avons la fonction de vraisemblance qui suit pour les probabilités de transition.

$$L(N; P) = \prod_{t \in \Xi^x} \prod_{v \in \Psi} \prod_{s \in \Theta} P(s|t)^{N(t, Z, v)}; \quad (2)$$

Dans (2), les indices exécutés sont t, Z, v et s . Si une réponse est fournie à chaque question, alors Ψ est le seul cas de Z pour lequel $N(t, Z, v) \neq 0$, pour certaines valeurs de t et v . Dans ces conditions, (2) est analogue à la vraisemblance de la probabilité de transition d'une chaîne markovienne de premier ordre (Bishop, Fienberg et Holland 1975, page 263). En général, nous modélisons Θ_p sous forme d'un sous-espace log-linéaire. À cette fin, il est plus pratique de se servir d'une expression équivalente de (2) dont la représentation algébrique est plus simple. Nous introduisons le paramètre nuisible $U = [U(t)]$, où U est un vecteur de probabilités, c'est-à-dire $\sum_{t \in \Xi^x} U(t) = 1$, et $0 < U(t) < 1$, pour tous les $t \in \Xi^x$. U représente les prévalences des états des plus proches voisins à réponse complète. Posons $\tilde{Q}(s, t) = U(t) \times P(s|t)$, et $\tilde{Q} = [\tilde{Q}(s, t)]$. Alors, \tilde{Q} est un vecteur de probabilités comptant $K \times F$ composantes classées lexico-graphiquement en fonction de t et de s . Nous définissons Θ , l'espace paramètre de \tilde{Q} , sous forme de modèle hiérarchique log-linéaire (Agerst 1990, page 143; Bishop, Fienberg et Holland 1975, page 67). Alors, si nous concevons Θ de sorte qu'il contienne les interactions de tous les ordres entre les variables comprises dans Ξ , (2) est équivalente à la fonction de vraisemblance suivante en ce qui concerne \tilde{Q} .

5. CALCUL DE L'EMV ET ÉTABLISSEMENT DE MESURES DE L'ERREUR DUE À LA NON-RÉPONSE

Autrement dit, si Θ possède l'architecture décrite plus haut, un choix spécifique pour Θ définit sans ambiguïté Θ_p dans (2), et, puisque les valeurs pour les questions correspondantes au voisin le plus proche à réponse complète sont toujours déclarées, la factorisation $L(N; P) = L^*(N; \tilde{Q}) \times R(N; U)$ est valable pour certaines valeurs de $R(\cdot)$. (3) est plus facile à manipuler que (2), puisqu'elle correspond à la vraisemblance des probabilités de cellules associées à un tableau de contingence classifié partiellement (Little et Rubin 1987, page 181). Sous des contraintes peu sévères appliquées au mécanisme de non-réponse (par exemple, probabilité strictement positive et constante pour chaque configuration de réponse (Thibaudau 1988)), les vraisemblances dans (2) et (3), sont identifiables et asymptotiquement unimodales. En théorie, la multimodalité est possible pour des échantillons finis, mais elle ne semble pas se produire dans les cas étudiés ici, où la proportion de questions sans réponse est faible.

$$L^*(N; \tilde{Q}) = \prod_{t \in \Xi^x} \prod_{v \in \Psi} \prod_{s \in \Theta} \tilde{Q}(s, t)^{N(t, Z, v)}; \quad (3)$$

À la présente section, nous rappelons comment calculer P , l'EMV de P , et nous établissons des mesures des erreurs pour $A(f)$ et un autre prédicteur $\hat{s}(s)$, que nous appelons « l'EMV » de $\hat{s}(s)$, qui est le dénombrement réel de chefs de ménage dans l'état s au niveau du secteur de recensement. Une mesure de l'erreur pour $\hat{s}(s)$ sera utile à la section 7 où nous comparons les résultats d'imputation obtenus au moyen de $A(f)$ et de la HDSPH. Pour calculer P , nous maximisons (3), par rapport à \tilde{Q} , au moyen de l'algorithme EM. Étant donné la factorisation décrite à la section 4, ce maximum donne aussi P .

Pour établir des mesures de l'erreur de prédiction de $\hat{s}(s)$ pour un s donné, considérons tous les triplets de forme (t, Z, v) dans (1) que l'on observe dans l'échantillon (c'est-à-dire le secteur de recensement) pour lesquels il est non-réponse partielle, qu'un ou plusieurs chefs de ménage correspondent à un tel triplet soit dans l'état s . Représentons par $\Lambda(s)$ le nombre de ces triplets. Nous donnons à ces triplets l'indice $\lambda = 1, \dots, \Lambda(s)$. Représentons par $\delta(\lambda)$ le nombre de chefs de ménage correspondant au triplet λ , et par $p_\lambda(s)$, la probabilité qu'un de ces chefs de ménage soit effectivement dans l'état s , où $p_\lambda(s)$ est tiré de P . Soit $\Delta(s, \lambda)$, le nombre inconnu de chef de ménage qui sont effectivement dans l'état s parmi les $\delta(\lambda)$ candidats. D'après notre modèle, nous avons $S(s) = S^{obs}(s) + \sum_{\lambda=1}^{\Lambda(s)} \Delta(s, \lambda)$, où $S^{obs}(s)$ est le nombre de chefs de ménage

Catégorie race-mode d'occupation du chef de ménage

	Nombre de chefs de ménage dans la strate à position 3									
	Blanc	Blanc	Noir	Noir	Asiatique	Asiatique	Autre	Autre	loc.	loc.
	prop.	loc.	prop.	loc.	prop.	loc.	prop.	loc.	prop.	loc.
Taux de propriété des plus proches voisins blancs	0,556	0,564	0,562	0,299	0,561	0,287	0,540	0,163		
Taux de propriété des plus proches voisins noirs	0,379	0,189	0,427	0,211	0,443	0,202	0,471	0,158		
Taux de propriété des plus proches voisins asiatiques	0,589	0,332	0,667	0,320	0,668	0,262	0,535	0,302		
Taux de propriété des autres plus proches voisins	0,423	0,251	0,497	0,237	0,595	0,177	0,463	0,152		

Racé du plus proche voisin

Race du plus proche voisin	Blanc	Noir	Asiatique	Autre
Taux de propriété des chefs de ménages blancs	0,415	0,358	0,384	0,337
Taux de propriété des chefs de ménages noirs	0,257	0,264	0,304	0,267
Taux de propriété des chefs de ménages asiatiques	0,441	0,441	0,400	0,360
Taux de propriété des chefs de ménages autres	0,309	0,297	0,337	0,234

TRANSITIONS DEMOGRAPHIQUES

Besag (1974) décrit l'application de la méthode des probabilités conditionnelles aux processus spatiaux. Cette méthode fournit un cadre pour la modélisation probabiliste des valeurs des « emplacements » en fonction des valeurs de leurs « voisins » pour construire un processus spatial. Besag (1974) propose aussi de faire une approximation unilatérale pour simplifier cette construction. Alors, la valeur de chaque emplacement dépend uniquement d'un nombre fini de « prédécesseurs ». Cette démarche est naturelle ici, puisque f fournit un ordonnancement unilatéral des chefs de ménage qui jouent tour à tour les rôles d'emplacement et de prédécesseur. Plus précisément, nous construisons un processus de premier ordre où chaque chef de ménage est un emplacement, et où le voisin le plus proche à réponse complète est son unique prédécesseur. Dans ces conditions, la valeur d'un emplacement est l'état d'un chef de ménage, que nous définissons brièvement. Nous définissons la probabilité conditionnelle pour la valeur d'un emplacement, étant donné celle de son prédécesseur, comme étant la probabilité de transition de l'état du voisin le plus proche à l'état du chef de ménage. Notre maximum de vraisemblance (EMV) des probabilités de transition au niveau du secteur de recensement. À la présente section, nous décrivons la méthode d'imputation, et à la suite, nous introduisons une fonction de vraisemblance pour les variables spatiales.

$$P(s|t, Z, v) = \frac{\sum_{v \in \sigma(\Psi, Z)} P(s|t)d(t|\Psi, Z, v)}{d(t|\Psi, Z, v)}; s \in \sigma(\Psi, Z, v). \quad (1)$$

concorde avec v sur les variables dans Z . Définissons sous-ensemble contenant toutes les valeurs de s , tel que s

variance, alors que, dans la même situation, l'estimateur jackknife (Rao et Shao 1992) ne produit pas de biais. Meng (1994) soutient que l'exemple de Fay a pour origine une mauvaise communication entre un imputeur qui dispose de renseignements spécifiques sur un modèle et un analyste qui ne connaît que le processus d'estimation. Dans les mots de Meng, ces situations sont incompatibles. Bien que l'exigence d'une coordination entre l'imputeur et l'analyste soit restrictive, l'imputation fondée sur l'échangeabilité présente aussi des défauts dangereux, comme nous le montrons à la section 2. En outre, la méthode bayésienne permet de procéder à l'approximation asymptotique des mesures de l'erreur grâce à des algorithmes mécaniques, alors qu'une méthode strictement fréquentiste pourrait nécessiter des développements fastidieux, comme nous le montrons à la section 5.

Notre objectif est de présenter $A(f)$ et de montrer ses avantages comparativement à la méthode HDSH, en nous servant de la répétition générale du recensement à Sacramento comme exemple. Dans ce cas particulier, f contient des enregistrements pour les 138 271 chefs de ménage de dénombrements physiologiques (Kostanich 1999), dont 90 156 ont retourné le questionnaire de recensement par la poste ou ont reçu la visite d'un recenseur lors d'une première tentative, et 48 115 ont été sélectionnés dans un échantillon. Nous appliquons notre méthode au niveau du secteur de recensement, c'est-à-dire une unité géographique liée contenant, en moyenne, 1 300 chefs de ménage dans f . La présentation de l'article est la suivante. À la section 2, nous illustrons les difficultés que pose la conception d'une méthode HDSH assurant l'échangeabilité. À la section 3, nous définissons $A(f)$, et à la section 4, nous présentons une fonction de vraisemblance pour les paramètres du modèle. À la section 5, nous montrons comment appliquer $A(f)$ et nous calculons une mesure de l'erreur pour procéder à des comparaisons avec la HDSH. À la section 6, nous présentons et justifions le modèle de base pour la répétition générale et à la section 7, nous donnons les résultats pour $A(f)$ ainsi que pour la HDSH dans ce cas. À la section 8, nous résumons les différences et faisons des recommandations.

2. ÉVALUATION DE L'ÉCHANGEABILITÉ POUR UNE SUBDIVISION PAR DOMAINE D'ÉTUDE

Nous illustrons les difficultés inhérentes à la conception d'une méthode HDSH qui préserve l'échangeabilité entre domaines d'étude (Cochran 1977, page 34) à l'aide d'un exemple où le mode d'occupation du logement (propriété) est la mesure étudiée et où les domaines d'étude pertinents sont définis par la race. Pour imputer les valeurs pour le mode d'occupation du logement, le Censur Bureau utilise la variable classe « type de ménage » pour stratifier f .

Nous examinons la strate a posteriori comprenant tous les chefs de ménage sans conjoint ou conjointe et vivant dans un ménage comptant trois personnes ou plus. Nous appelons strate a posteriori 3. Pour les besoins de l'exemple, nous avons supprimé de f tous les chefs de ménage qui n'ont pas déclaré leur mode d'occupation du logement, et chaque plus proche voisin ne se rapporte qu'à un seul chef de ménage. Le tableau 1 donne la fréquence des chefs de ménage pour huit catégories race-mode d'occupation du logement exhaustives pour la strate a posteriori 3. Il donne aussi le taux de propriété pour ces plus proches voisins, classifiés selon la race et les mêmes huit catégories race-mode d'occupation du logement que pour les chefs de ménage correspondants. Nous constatons qu'en moyenne, lorsqu'un chef de ménage appartient à la catégorie Noir-propriétaire ou Noir-locataire, son plus proche voisin est au moins 25 % plus susceptible d'être propriétaire s'il est blanc que s'il est noir. Il est tentant de s'appuyer sur des différences géographiques pour expliquer ce taux différentiel. Cependant, le tableau 2, qui donne les taux de propriété pour les chefs de ménage de la strate a posteriori 3, classés selon leur race et celle de leur plus proche voisin, montre qu'en fait, le taux de propriété est légèrement plus faible pour les Noirs dont le plus proche voisin est blanc que pour ceux dont le plus proche voisin est noir. Autrement dit, si la probabilité de ne pas déclarer le mode d'occupation du logement est constante chez les Noirs dans l'ensemble, alors imputer le mode d'occupation déclaré par le plus proche voisin donne lieu à une surestimation de la propriété chez les Noirs dans la strate a posteriori 3.

Les différences distributionnelles entre les chefs de ménage et leurs plus proches voisins reflètent un manque d'échangeabilité. Un test de McNemar mène au rejet formel de l'hypothèse d'échangeabilité. Dans notre exemple, 1 784 chefs de ménage noirs ont un plus proche voisin blanc. Dans 1 187 cas, le mode d'occupation du logement est le même. Pour 396 des 597 cas où le mode d'occupation n'est pas le même, le propriétaire est blanc. Aux termes de l'hypothèse d'échangeabilité, parmi les propriétaires la proportion de blancs parmi les propriétaires est supérieure de huit écarts-types à la moitié. Cet exemple illustre les difficultés que pose la conception d'une méthode HDPV valide respectant le concept d'échangeabilité. À la section suivante, nous présentons notre méthode d'imputation, conçue pour faire face à ce type de situation.

néage qui les a déclarées et qui se trouve dans la même strate à posteriori (Treat 1994). L'ordonnement tridimensionnel des volumes géographiques. Dans le cas de la HDSH, l'intention est de définir des « plus proches voisins » qui sont proches géographiquement ainsi qu « en nature ». Dans la suite de l'article, nous continuons d'utiliser l'expression chef de ménage, mais sa signification peut s'étendre à une unité d'échantillonnage géographique.

La conception de la HDSH convient bien à l'imputation pour la correction de la non-réponse partielle dans le cas de populations regroupées géographiquement par domaine. Dans ce cas, le besoin de variables classes est limité. Toutefois, des difficultés surviennent lorsque les limites géographiques, entre les domaines, deviennent floues. Concevoir une méthode HDSH dont le pouvoir discriminatoire est bon dans ces conditions revient à trouver le juste équilibre entre la spécification d'un nombre suffisant de variables classes pour tenir compte des hétérogénéités entre domaines et la sous-spécification d'un trop grand nombre de ces variables, qui pourraient produire des strates à positions définies si étroitement en ce qui concerne le domaine qu'elles ne

Le fait que la composition démographique de la population peut varier lorsque l'emplacement géographique change complique la situation et, par conséquent, il pourrait être nécessaire de réviser un scénario particulier de HDSh en fonction du lieu géographique. Étant donné ces difficultés, $A(f)$ est une innovation en ce sens que, au lieu de rechercher un plus proche voisin idéal, elle produit des imputations grâce à une simulation basée sur un modèle qui intègre les renseignements sur la géographie locale, ainsi que les subdivisions par domaine. $A(f)$ intègre les deux types d'information par calage des paramètres d'un modèle log-linéaire, d'après la force des corrélations entre les covariables et les variables susceptibles de faire l'objet d'une imputation. Notre stratégie d'estimation des paramètres est la même que celle de Zanutto et Zaslavsky (1995a, b). Cependant, parce qu'ils disposent d'un échantillon représentatif de non-répondants complets, ces auteurs peuvent estimer les probabilités d'imputation par application d'un algorithme EM à une étape (Dempster, Laird et Rubin 1977). Dans notre situation, nous ne supposons pas que nous disposons d'un échantillon représentatif et nous appliquons l'algorithme EM complet. Implicitement, nous émettons l'hypothèse que les réponses « manquent au hasard » (MAR pour *missing at random*) (Little et Rubin 1987, page 16).

Pour analyser les résultats obtenus au moyen de $A(f)$, nous comparons à ceux de la HDSh, nous déterminons les mesures d'erreur associées à $A(f)$, d'après des approximations calculées au moyen d'un algorithme bayésien introductif pour la première fois par Gelman et Rubin (1991). Il existe des objections fondamentales à l'application de méthodes bayésiennes. Fay (1992) montre que l'estimation de la variance fondée sur des imputations multiples (Rubin 1996) peut produire des estimations gonflées de

méthodes hot-deck, à savoir la méthode hot-deck par le plus proche voisin (HDPV), Chen et Shao (1997, 2000) donnent une définition abstraite de la HDPV en fonction d'une mesure de la proximité $| \cdot |$, fondée sur une covariable x . Selon cette méthode DDPV, un "donneur" x correspond à x_i importe quelle unité telle que $|x_i - x_j|$ soit minimale, où x_j correspond à l'unité bénéficiaire (receveur) et x_i correspond au fournisseur des données imputées (donneur).

variable x appropriée, nous retrouvons la HDS pure et la HDCH pure sous forme de cas particuliers de la HDPV. La HDCH pure consiste à imputer la valeur d'une réponse au receveur en remplaçant celle-ci par la valeur correspondante provenant de l'unité la plus proche pour laquelle une réponse a été reçue, selon l'ordre établi dans f . La HDCH pure se fonde uniquement sur la valeur des variables HDCH pure et celle à laquelle appartient le receveur.

Fay (1999), et Fay et Town (1998) proposent le concept d'échangeabilité pour valider la HDPV. Pour des données catégoriques, deux unités de f sont échangeables si elles sont non corrélées et identiquement distribuées, étant donné l'information dont on dispose avant l'imputation. Les HDS pure, le concept signifie que deux unités contigües dans f sont échangeables. Dans celui de la HDCF pure, il signifie que les unités dont les valeurs des variables latentes sont les mêmes sont échangeables sans égard à leur position dans f . Nous définissons une troisième version de la HDPV, que nous appelons hot deck séquentielle hybride (HDSH). Dans le cas de cette dernière, pour qu'il y ait échangeabilité, il faut que les conditions de proximité soient remplies tant pour l'ordonnement du fichier f que pour les variables latentes.

Sans indication contraire, nous utilisons l'expression « plus proche voisin » au sens abstrait de la HDPV. Nous utilisons les expressions « voisin le plus proche » pour désigner le plus proche dans le cas de la HDS pure et « voisin le plus proche à réponse complète » pour désigner l'unité d'échantillonnage sans non-réponse partielle la plus proche selon l'ordonnement de f . Dans le cas de la répartition générale du recensement à sacramento, le Census Bureau utilise une méthode HDSH pour estimer le nombre de chefs de ménage, selon le mode d'occupation du logement, la race, l'origine hispanique) et le sexe. Le chef de ménage, qui est habituellement un adulte et dont le nombre est de 1 par unité de logement, est repéré d'après les âges, les items et l'ordre d'énumération des personnes sur le questionnaire de recensement. La HDSH consiste à remplacer les valeurs manquantes par les valeurs fournies pour les questions correspondantes par le dernier chef de

Imputation pour la non-réponse partielle basée sur un modèle explicite pour les catégories démographiques

YVES THIRAUDEAU¹

RÉSUMÉ

Nous proposons une méthode d'imputation pour corriger la non-réponse partielle applicable aux données catégoriques fondée sur un estimateur du maximum de vraisemblance (EMV) établi d'après un modèle à probabilités conditionnelles (Besag 1974). Nous définissons aussi une mesure de l'erreur due à la non-réponse partielle utile pour évaluer le biais comparativement à celui produit par d'autres méthodes d'imputation. Pour calculer cette mesure, nous procédons à un ajustement proportionnel itératif bayésien (Gelman et Rubin 1991; Schafer 1997). Nous appliquons notre méthode d'imputation à la répétition générale de 1998 du Recensement de 2000 à Sacramento et nous utilisons la mesure de l'erreur pour comparer l'imputation pour la non-réponse partielle par notre méthode et selon une version de la méthode hot-deck par le plus proche voisin (Fay 1999; Chen et Shao 1997, 2000) à des niveaux agrégés. Nos résultats donnent à penser que notre méthode protège mieux que la méthode hot-deck contre le biais d'imputation dû à l'hétérogénéité des domaines d'étude.

MOTS CLÉS : Plus proche voisin; méthode des probabilités conditionnelles; ajustement itératif proportionnel bayésien.

1. INTRODUCTION ET CONTEXTE

Représentons par S un dénombrement démographique par catégorie demandé dans le cadre d'un recensement ou nécessaire pour calculer une statistique d'enquête, et supposons que l'on peut calculer S d'après les enregistrements d'un fichier d'enquête f , lorsque les enregistrements sont complets. En outre, supposons que f est ordonné de telle façon que la proximité selon l'ordonnement de f correspond à la proximité géographique. Considérons la situation où f comprend des enregistrements pour lesquels les réponses à certaines questions manquent. Nous proposons d'estimer S au moyen de $d(A(f))$, où $A(f)$ est une méthode d'imputation qui produit un fichier de données d'enquête complet, et où $d(\cdot)$ estime S par remplacement des réponses manquantes par les valeurs correspondantes imputées au moyen de $A(f)$. $A(f)$ est basée sur une fonction de vraisemblance qui modélise les transitions entre deux enregistrements voisins dans f , et sur des associations entre les questions auxquelles il faut imputer une réponse et les domaines d'étude pertinents (Cochran 1977, page 34) définis par des subdivisions de la population. $A(f)$ a pour but de remplacer avantageusement la méthode hot-deck séquentielle couramment utilisée (Kovar et Whitridge 1995), qui est une version de la méthode hot-deck par le plus proche voisin (Fay 1999; Chen et Shao 1997, 2000) visant à réduire au minimum la distance géographique entre une unité pour laquelle la réponse à certaines questions n'est pas déclarée et un enregistrement donneur approprié pour l'imputation, tout en assurant l'homogénéité distributionnelle des réponses observées et imputées pour chaque domaine étudié. Si les

domaines d'une même subdivision ont tendance à ne pas se chevaucher géographiquement, emprunt, aux fins d'imputation, de réponses à un enregistrement voisin préserve l'homogénéité. Par contre, si les petits domaines ont tendance à être dispersés dans de grands domaines, un dilemme se pose au méthodologiste. Dans ces circonstances, il doit faire un choix entre les règles hot-deck qui mènent à l'emprunt de réponses à des unités géographiquement proches, susceptibles de donner lieu à des biais d'imputation reflétant l'hétérogénéité locale entre les domaines, et les règles particulières à un domaine, qui assurent l'homogénéité distributionnelle par domaine, mais ne réduisent pas au minimum la distance géographique. $A(f)$ est une solution de recherche conçue pour maintenir l'intégrité du domaine, tout en simulant le profil distributionnel d'un enregistrement donneur ayant certaines caractéristiques en commun avec un enregistrement géographique. $A(f)$ est une solution de recherche conçue pour maintenir la réponse partielle et de leurs principes opératoires, afin de pouvoir les comparer correctement à $A(f)$ plus loin. Nous donnons aussi des précisions sur la répétition générale du Recensement de 2000 à Sacramento qui nous sert de banc d'essai dans tout l'article.

Fay (1999) et Sande (1981) considèrent la méthode hot-deck séquentielle (HDS) comme étant la première catégorie de méthodes hot-deck, que nous appellerons HDS « pure ». Ils ajoutent une deuxième catégorie, la méthode hot-deck à cellules fixes (HDCF), que nous appellerons HDCF pure. Fay définit une troisième catégorie de

Nous convenons tous que le souci de la qualité représente un aspect fondamental de notre activité. Le présent document vise à montrer qu'il y a de nombreux aspects à la qualité. C'est le même message que livrent nettement les cadres de qualité dressés par d'autres organismes, comme le FMI, Statistique Canada et Statistique Suède. La conséquence en est qu'un organisme de qualité dépend des gestes de tous ses employés, car tous peuvent agir sur la qualité d'une manière ou d'une autre. On ne peut laisser la responsabilité de la qualité uniquement à un groupe de travail. Il n'y aura donc qualité que s'il régit une authentique culture de la qualité dans un organisme. L'article décrit comment nous réalisons la chose à l'ABS. Il importe néanmoins que quelqu'un devienne la « conscience » d'un organisme en matière de qualité. C'est la méthodologie, dont le chef fait aujourd'hui partie de l'équipe de direction, ce qui facilite la communication des grands messages à la haute direction. Les membres de la division attirent notamment l'attention sur les risques les plus importants ou les comportements contraires aux visées de l'organisme dans le domaine de la qualité.

5. CONCLUSION

REMERCIEMENTS

La section 2 du présent document s'inspire en partie d'un document produit par Frank Yu (ABS) pour le 9^e congrès des organismes statistiques d'Asie de l'Est.

BIBLIOGRAPHIE

- ALLEN, B. (2001). *Qualifying Quality – Issues of Presentation and Education. Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistique Canada.
- BRACKSTONE, G. (1999). *La gestion de la qualité des données dans un bureau de statistique. Techniques d'enquête*. 25, 157-171.
- CARSON, C. (2000). *Towards a framework for assessing data quality. The Proceedings of the Statistical Quality Seminar, Jeju*. Korean National Statistical Office and International Monetary Fund.
- FELLEGI, I.P. (1996). *Characteristics of an effective statistical system. Revue Internationale de Statistique*. 64, 165-197.
- LBE, G., et ALLEN, B. (2001). *Educated Use of Information about Quality. Bulletin of the International Statistical Institute*, 53^e Session, Séoul, Corée.

- Comme nous avons résolulement entrepris de normaliser nos méthodes et nos systèmes, nous dépendons moins des connaissances locales.
- Dans certains postes clés (direction du service de la comptabilité nationale, par exemple), nous assurons un jumelage avant la retraite du titulaire.

Jusqu'ici, nous avons bien ménagé les transitions. Nous avons pu pourvoir aux postes vacants à la haute direction tout en rajeunissant l'organisme par l'avancement d'employés ayant des idées neuves à proposer. Il est nécessaire que nous continuions à agir avec adresse

4.7 Application des normes internationales

Notre position de départ est que, là où il existe des normes internationales, nous nous devons de les appliquer. Tel n'a pas toujours été le cas. Ainsi, bien que notre classification des industries soit plus ou moins fondée sur la CITI et qu'une table de concordance ait été dressée avec cette classification, notre nomenclature demeure largement un produit du cru qui traduit les intérêts particuliers de l'Australie et de la Nouvelle-Zélande. Nous avons accepté d'appliquer la version 2007 de la CITI, du moins à ses deux niveaux supérieurs, en prévoyant des dérogations aux niveaux inférieurs seulement si les circonstances l'exigent. Souvent, des pressions s'exercent sur nous pour que nous nous écartions des normes internationales. Quelquefois, on veut que la situation australienne paraisse meilleure. Dans d'autres cas, comme celui de la définition du chômage de l'OIT, ces pressions tiennent au fait que la définition internationale ne semble pas rendre compte de la situation concrète de l'Australie. Nous résistons à ces pressions, mais il importe que, pour justifier notre position, nous puissions en référer à une norme internationale bien décrite. Il reste que, si nous dérogeons exceptionnellement à la norme internationale, il nous faut bien préciser les circonstances et exposer clairement nos motifs. Là où nous éprouvons le besoin de disposer de renseignements qui ne sont pas fondés sur la norme internationale, la règle est de publier des statistiques sur les deux bases. Les chiffres principaux devraient être conformes à la norme internationale, car on compare de plus en plus la situation de l'Australie à celle des autres pays et la comparabilité est donc chose importante. C'est la position que nous adoptons pour répondre à la demande de données sur le sous-emploi dans l'économie et combattre les critiques dont est l'objet la définition du chômage de l'OIT.

Il y a ainsi une tension à gérer, mais si pour nous les comparaisons internationales sont quelque chose de sérieux, il est impératif que la norme internationale nous serve principalement de guide au moment d'élaborer des concepts, des sources et des méthodes pour l'Australie. Par conséquent, nous jugeons prioritaire d'apporter une bonne contribution à l'élaboration et à la révision des normes internationales.

4.6 Gestion du transfert de connaissances et des compétences

Des détails supplémentaires figurent dans Allen (2001).

- établissement de quatre prototypes d'instruments destinés à aider les usagers à comprendre la qualité de statistiques particulières : « sommaires des questions de qualité », « mesures de qualité », « exactitude des données » et « accès intégré aux données et aux métadonnées ».

En revanche, nous perdons de l'expérience et du savoir-faire. Il faut savoir soigneusement gérer les deux côtés de l'équation. Voici notre stratégie :

- Nous avons conçu des programmes spéciaux pour les employés promoteurs. Plus précisément, ceux-ci suivent un programme de perfectionnement en leadership et en gestion qui est adapté aux besoins de l'ABS. Ce sont les cadres supérieurs qui choisissent les candidats à de tels programmes. Il est impossible à quelqu'un de choisir lui-même de participer à un programme de perfectionnement. Ajoutons que, une fois achevé ce programme, les gens peuvent s'attendre à recevoir une affectation spéciale ou une désignation par roulement à un nouveau poste. Le principe fondamental est que le meilleur moyen d'apprentissage est la riche diversité des expériences vécues au travail. Une très forte proportion des gens qui ont récemment été promus à des postes de direction ont participé à des programmes de perfectionnement similaires. Cela nous a aidés jusqu'à présent à bien combler les vides créés par les passages à la retraite qui se sont multipliés.
- Nous gardons des liens avec les retraités de l'ABS par diverses voies officielles ou officielles (rencontres sociales, inscription sur la liste de diffusion de « ABS News », etc.). Nous pouvons faire appel à leur savoir au besoin.
- Comme nous mettons plus l'accent sur la gestion des connaissances grâce aux moyens que nous offre notre collectif Lotus Notes, les parties essentielles de notre activité sont bien décrites et cette documentation est facile à consulter.

tout aussi embarrassantes pour nous que celles des produits sur papier.

Nos méthodes d'assurance de la qualité des produits électroniques ne sont pas aussi raffinées, mais elles évoluent. Voici les grandes mesures prises dans ce domaine :

- Notre entrepôt d'information soutient le stockage de tous les objets ayant à voir avec la diffusion de tel ou tel ensemble de statistiques, ce qui comprend les cubes d'information et les métadonnées.
- Les secteurs statistiques sont priés d'approuver chaque objet; ils conçoivent individuellement leurs propres techniques d'assurance de la qualité (tout en échangeant des idées sur les meilleures pratiques).
- On a mis en place un système d'édition pour la diffusion simultanée de tous les produits; si ceux-ci émanent du même ensemble d'objets, les risques d'incohérence de produits seront moindres.

4.5 Communication de statistiques dans Internet

En dernière analyse, l'utilisateur ne peut juger que de l'adéquation d'un produit statistique à ses propres usages. Ceux-ci varient, bien sûr, et ce qui convient à un usage pourrait ne pas convenir à un autre. Nous avons l'obligation de fournir une diversité d'information à l'appui de ce que nous produisons comme statistiques, y compris des renseignements sur la qualité, de sorte que l'utilisateur puisse porter son propre jugement sur l'adaptation à l'usage. Il existe un certain nombre de pratiques éprouvées en matière de déclaration de la qualité des statistiques. Ces activités font aujourd'hui partie intégrante des pratiques de diffusion adoptées :

- Il y a des publications sur les concepts, les sources et les méthodes qui décrivent les méthodes d'élaboration des principaux produits statistiques. On peut les trouver à notre site Web, ainsi que sur d'autres supports.
- Il y a une diversité de documents d'information et de travail et d'articles de fond dans les publications qui attirent l'attention sur les particularités de produits ou sur les modifications apportées à ces mêmes méthodes.
- On applique une politique d'« absence de surprises » si les méthodes d'élaboration des séries statistiques subissent des modifications importantes. Il n'y a pas que les documents d'information ou autres en cas de refonte de séries statistiques, puisque nous menons auprès des grands utilisateurs un programme de séminaires et de discussions bilatérales permettant d'expliquer les changements et leurs motifs.

- Il y a des indications sur les méthodes employées dans toutes nos publications. L'ordre et les éléments matériels de présentation de cette information sont conformes à des normes convenues. Nous avons mis celles-ci au point à la suite de recherches faites par un conseiller en communications sur la façon dont nos utilisateurs exploitent le contenu de nos publications statistiques.
- La partie analytique de nos publications explique entre autres les mouvements amples ou inusités de nos séries statistiques. Souvent, ces explications reposent sur des indications que le personnel de l'ABS est seul à obtenir par ses contacts avec les enquêtes ou par sa profonde connaissance des méthodes d'élaboration de statistiques. Nos groupes d'utilisateurs nous ont dit que c'était là une des formes d'analyse les plus utiles auxquelles nous puissions nous livrer.

À notre avis, nos principaux utilisateurs comprennent assez bien la qualité des statistiques qu'ils utilisent, mais la plus grande dépendance à l'égard de la diffusion électronique est source de nouveaux défis. En un sens, cette évolution est une occasion en or de livrer sur les questions de qualité une information diverse qui puisse facilement être consultée en quelques « clics » bien appliqués. Mais comme l'information sur la qualité des statistiques n'est pas aussi évidente qu'elle peut l'être dans les publications sur papier, les utilisateurs peuvent plus facilement éviter les messages clés que nous essayons de leur transmettre. Le véritable défi pour nous est de concevoir des méthodes de présentation des questions de qualité de sorte qu'il soit difficile à l'utilisateur de ne pas prendre connaissance des grands messages que nous entendons livrer.

- Une manière d'agir peut être d'y aller de messages séparés qui attirent l'attention sur des éléments d'information particuliers que l'on veut voir l'utilisateur recevoir sur les questions de qualité. Ces messages pourraient automatiquement être activés lors de l'accès à des séries statistiques déterminées ou faire l'objet d'une communication distincte par courrier électronique. On doit à cet égard étudier les moyens les plus efficaces.
- Lee et Allen (2001) ont décrit certaines des recherches effectuées à ce jour dans ce domaine. Nous en sommes encore à un stade exploratoire. Voici ce sur quoi porte notre investigation :
- tests de facilité d'utilisation visant à déterminer ce que les utilisateurs préfèrent comme voie d'accès à l'information sur la qualité;
- leadership et élaboration de programmes d'éducation de l'information sur la qualité (une version expérimentale est maintenant disponible);

croyons cependant pas que les méthodes de mesure statistique soient un facteur qui intervient beaucoup, puisque, le plus souvent, leur évolution s'est traduite par des améliorations, bien que les perceptions puissent être différentes. Ainsi, les grandes séries de la comptabilité nationale sont bien moins instables qu'il y a 10 à 15 ans et, pourtant, certains utilisateurs voient la chose bien autrement.

Nous recevons aussi plus de critiques au sujet d'hexactitudes relevées dans les données très fines (tableaux du recensement de la population, par exemple). Là encore, ce n'est pas que la qualité se dégrade, c'est que les attentes se sont élevées.

Il nous faut accepter que la « barre soit plus haute » et faire tout notre possible pour porter la qualité au niveau de ces attentes grandissantes. Cela n'est pas toujours chose faisable, bien sûr; aussi une gestion des attentes importe-t-elle. On peut ainsi :

- donner de bonnes explications des forces et des faiblesses d'ensembles de données particuliers;
- parler aux principaux utilisateurs des forces et des faiblesses des séries de données partout où on peut le faire;
- réagir aux critiques éclairées (rechercher des collaborations pour l'amélioration de la qualité; ainsi, dans nos statistiques détaillées sur le commerce extérieur, nous sollicitons ouvertement une rétroaction des utilisateurs au sujet de la qualité des statistiques produites);
- expliquer le plus possible les statistiques qui peuvent paraître injustes ou inattendues.

4.3 Amélioration des méthodes d'exploitation statistique

À l'instar de plusieurs organismes statistiques, l'ABS se demande comment il pourrait recourir aux nouvelles technologies et exploiter d'autres possibilités comme celle d'une utilisation accrue des données fiscales afin de rendre plus efficaces ses procédés d'exploitation en statistique des entreprises.

Il scrute aussi ses méthodes d'enquête auprès des entreprises, compte tenu du recours accru à l'interview assistée par ordinateur (IAO) en particulier. Toutefois, cette section se concentrera sur les changements apportés aux méthodes de gestion de la statistique des entreprises pour décrire le défi à relever en matière de qualité.

L'ABS a chargé une équipe d'examiner les diverses possibilités. À la suite des travaux de cette équipe, un certain nombre de changements importants ont été approuvés dans le cadre d'un programme d'innovation en statistiques des entreprises. Nous cherchons à réviser nos processus opérationnels et à nous doter pour au moins 10 ans de processus qui garantiront un rendement appréciable de nos investissements nécessaires de mise en place. Nous allons :

- étendre les responsabilités du service du registre des entreprises pour recueillir et stocker des données fiscales en liaison directe avec le registre des entreprises grâce au numéro d'entreprise australienne, lequel est maintenant attribué par le canal du régime d'inscription fiscale des entreprises et se trouve dans la plupart des bases de données transactionnelles de ces mêmes données; les données seront stockées de manière appropriée; les données exploitées par les divers secteurs statistiques de l'organisme au moment d'assembler des données de statistiques à partir des données fiscales ou en combinaison avec les données d'enquête de l'ABS;
 - améliorer les méthodes utilisées pour traiter avec les entreprises visées par nos enquêtes, notamment leur donner une certaine latitude quant à la façon dont elles nous fournissent les données;
 - créer un entrepôt de données d'entrée où l'élément de raccordement des divers ensembles de données sera le numéro d'entreprise australienne;
 - mettre en place un environnement de traitement de données sur les entreprises axé sur l'entrepôt de données d'entrée;
 - centraliser davantage un certain nombre de fonctions d'élaboration de la statistique des entreprises.
- Nous pouvons discerner les avantages de cette évolution : production plus efficiente des statistiques sur les entreprises, meilleur recours aux données fiscales et à d'autres données administratives et bases de données qui soutiennent un plus large éventail d'analyses statistiques. On réduira toutefois ainsi le lien entre les secteurs qui produisent les statistiques et les sources de données d'entrée. Quelle en sera l'incidence sur la qualité? Quelles stratégies pouvons-nous appliquer pour diminuer cette incidence? Ce sont là d'importantes questions auxquelles nous devons répondre. Il s'agit du plus grand risque à gérer durant l'exécution du programme d'innovation en statistique des entreprises.

4.4 Assurance de la qualité des produits électroniques

L'ABS veille jalousement sur la qualité de ses produits sur papier et compte de nombreuses années d'expérience dans ce plan. Sa feuille de route parle d'elle-même dans le domaine et les techniques d'assurance de la qualité sont bien ancrées dans les modes d'exploitation. Il reste que de plus en plus d'utilisateurs reçoivent leurs données sur support électronique seulement et se fonderont sur ces produits pour faire leurs analyses et souvent prendre des décisions importantes en conséquence. Les erreurs qui risquent de se glisser dans les produits électroniques sont

4.1 Utilisation croissante de bases de données administratives ou transactionnelles

Nous nous servons depuis longtemps de bases de données administratives (c'est-à-dire registres d'état civil des naissances et des décès, données douanières sur le commerce, etc.) pour produire des statistiques officielles. Nous en avons employé d'autres pour établir des cadres de collecte statistique. Les questions à résoudre sont celles du nombre croissant de telles bases de données, de leur sous-utilisation à des fins statistiques et de l'exploitation des possibilités de coupler ces bases de données entre elles et aux ensembles de données de l'ABS au moyen d'un code d'identification commun (par exemple, le numéro d'entreprise australienne dans la statistique des entreprises).

Comme exemples de bases de données administratives devenues disponibles, on peut citer les bases étendues de la fiscalité des particuliers et des entreprises, les bases de données transactionnelles de l'assurance-maladie et les banques de données détaillées des régimes de soutien du revenu.

Les bases de données transactionnelles deviennent accessibles, mais elles ne se présentent pas sous une forme immédiatement exploitable. Celles qui présentent un intérêt particulier pour l'ABS sont les bases de données de lecture optique des magasins de détail et celles de transmission électronique de fonds au point de vente (c'est-à-dire les transferts électroniques de fonds entre clients et détaillants). Il y a des avantages à utiliser les bases de données administratives ou transactionnelles :

- on allège le coût d'information que l'on impose aux enquêtes;
- comme il s'agit souvent de « recensements », il est possible de produire des ensembles de données détaillées (selon les régions, par exemple);
- les bases ont souvent un caractère longitudinal (données fiscales, par exemple) qui favorise une analyse de ce genre;
- elles comportent souvent un code d'identification qui facilite l'analyse entre ensembles de données (par exemple, l'entreprise australienne, qui simplifie l'analyse entre les ensembles de données sur l'imposition des entreprises, les données douanières et les données des enquêtes de l'ABS).
- les sont parfois moins coûteuses que les ensembles de données recueillies de façon distincte.

Elles ont des défauts, bien sûr : définitions peut-être incohérentes par rapport aux concepts statistiques privés, souci moindre de la qualité des données d'entrée, information qui risque d'être dépassée, etc. La gestion de tout ce qui est protection des renseignements personnels est particulièrement importante. Bien que nés des motifs les plus honorables, qui servent les intérêts du public, les couplages de bases de données ont tout d'une question

4.2 Attentes grandissantes des utilisateurs

Les attentes des utilisateurs en matière de qualité évoluent. Elles sont bien plus élevées maintenant qu'il y a seulement cinq à dix ans. C'est une tendance qui devrait se maintenir. Avec les progrès de la mondialisation des marchés financiers, les grandes statistiques macro-économiques acquièrent une importance non seulement nationale mais internationale.

On a l'impression que les statistiques sont plus changeantes qu'avant. C'est que, dans certains cas, les phénomènes décrits sont devenus plus instables. Nous ne

Nous élaborons des protocoles de publication et de gestion des données puisées à des sources administratives. Il y a aussi la promotion et le soutien de bonnes pratiques de gestion des statistiques et des données.

Pour chaque domaine statistique, nous dressons, en collaboration avec d'autres intervenants, des plans de développement de l'information qui indiquent les secteurs de première importance et les activités qui rendront plus disponibles les données n'émanant pas directement de l'ABS; on y traite particulièrement des questions de gestion de la qualité.

Nous nous employons activement à promouvoir l'adoption de bonnes pratiques de gestion de l'information. Un projet d'investissement important vise à une plus grande utilisation des données fiscales pour une production statistique rentable.

Nous étudions des méthodes d'assurance de la qualité des ensembles de données très abondants mais imparfaits que l'on peut tirer des bases de données administratives et transactionnelles.

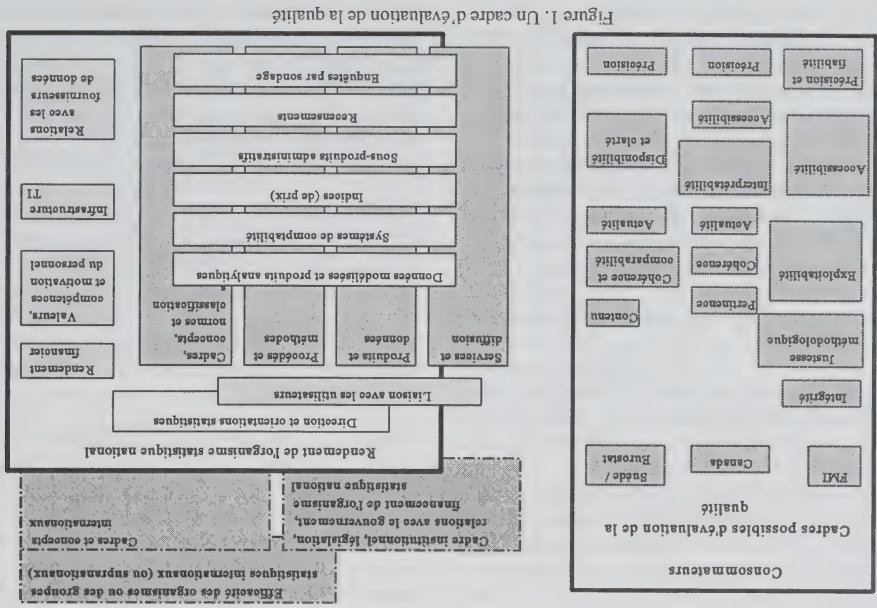
4. DEFIS COURANTS EN MATIERE DE QUALITE A L'ABS

Puisque les psychologues disent qu'il est difficile d'assimiler plus de sept points à la fois, la suite de cet exposé se bannera à traiter de sept grands défis que doit relever l'ABS en matière de qualité.

- (i) Utilisation croissante de bases de données administratives ou transactionnelles imparfaites pour l'élaboration des statistiques officielles;
- (ii) attentes grandissantes des utilisateurs qui font remonter la « barre » de la qualité;
- (iii) gestion de la tension entre l'amélioration des méthodes d'exploitation statistique (ce qui veut parfois dire que les responsables des « sorties statistiques » ne sont pas directement associés aux procédures relatives aux « entrées statistiques ») et le maintien ou l'amélioration de la qualité des produits statistiques;
- (iv) assurance de la qualité des produits électroniques;
- (v) présentation de statistiques sur Internet, avec le besoin d'éduquer la communauté de ses utilisateurs pour ce qui est de la qualité des statistiques officielles;
- (vi) gestion du transfert de connaissances et de compétences dans une situation où un grand nombre de membres vieillissants de la haute direction prendront leur retraite d'ici cinq ans;
- (vii) application de normes statistiques internationales à des fins de comparabilité là où ces normes pourraient ne pas le mieux convenir aux statistiques nationales.

- (ii) Il y a plusieurs groupes d'activités liés aux produits statistiques, depuis les « cadres, concepts, normes et classifications » jusqu'aux « services et diffusion ». Chacun a de l'importance en soi et comporte ses propres défis sur le plan de la qualité;
- (iii) Le rendement d'un organisme statistique national est extrêmement important pour l'image que ce dernier projette en matière de qualité, ainsi qu'il l'est reconnu dans la citation de la première page du présent document. Un certain nombre d'éléments de rendement sont indiqués à la figure 1. Tous sont importants. Un organisme statistique ne saurait être hautement performant si son rendement laisse à désirer pour un de ces éléments, sans oublier le rendement en gestion et dans les finances.
- (iv) Il y a d'autres éléments d'importance comme le cadre institutionnel (cadre législatif, par exemple).

Le but principal de la description qui précède est de souligner que la liste des défis de la qualité qui se présentent à un organisme statistique national est très longue. On doit s'atteler à toutes ces tâches d'une manière ou d'une autre, ce qui serait impossible sans une culture de la qualité, c'est-à-dire sans un souci de la qualité que doit cultiver tout le personnel. Il y a bien des « moments de vérité » où l'existence d'une culture de la qualité est véritablement mise en doute.



collecte statistique, la gestion de la qualité s'appuie sur des jeux d'indicateurs de rendement. On a conçu un ensemble type de mesures permettant une comparaison de qualité entre collectes de données. On met actuellement des outils au point pour intégrer les calculs relatifs à ces mesures aux activités normales d'enquête. L'entrepôt d'information sera le moyen de stocker et de présenter ces mesures. Les grands indicateurs retenus figurent aussi dans les rapports annuels que soumettent les différentes directions générales aux dirigeants de l'ABS.

Les mesures de la qualité intéressent les utilisateurs des statistiques. L'entrepôt d'information rendra les renseignements sur les questions de qualité plus accessibles aux usagers. Ajoutons que l'ABS juge hautement prioritaire d'aider l'utilisateur à comprendre la qualité des données et ses conséquences en ce qui le concerne et que, pour dynamiser cette compréhension, il s'est doté de stratégies nouvelles d'éducation de la clientèle. Comme le souligne Lee et Allen (2001), il y a fort à faire pour donner aux utilisateurs une meilleure compréhension du phénomène de la qualité.

Dans chaque enquête de l'ABS auprès des ménages, on prévoit aujourd'hui un programme d'évaluation de l'efficacité et de l'efficience de toutes les activités qu'elle comporte et du degré d'exploitation des données par les clients. Le centre d'information statistique examine toutes les enquêtes-entreprises de l'organisme. Grâce à ces initiatives, on s'assure que toutes les collectes font au moins l'objet d'une évaluation de base et de dégage des possibilités d'amélioration de la qualité et de l'efficience.

Non seulement l'ABS procède à des comparaisons internes de rendement entre ses propres secteurs de collecte, mais il a mis en place un réseau d'analyse comparative avec les organismes statistiques d'autres pays, ce maillage visant à des échanges d'information sur les plans, les procédés et les coûts des enquêtes. Cet exercice analytique oriente très utilement les efforts que déploie l'ABS en vue d'améliorer ses méthodes et ses produits.

2.6 Personnel compétent et motivé

L'ABS ne saurait fournir des statistiques de grande qualité à la communauté de ses utilisateurs s'il n'avait à son service des gens qui soutiennent les travaux statistiques de leurs compétences et de leurs efforts. Le personnel est chargé d'appliquer les stratégies que nous avons évoquées. Il doit faire preuve de professionnalisme et faire sien les tâches d'élaboration de nouvelles méthodes, d'amélioration continue de la qualité et de libre discussion des questions de méthodologie et de qualité.

Le personnel partage son temps et son énergie entre les travaux statistiques permanents et les tâches d'amélioration de la qualité. La règle dans l'organisme est d'intégrer autant que possible les tâches relatives à la qualité aux procédés et aux systèmes permanents. On y répète que la gestion de la qualité est une priorité de l'organisme et on veille à ce que des outils et des ressources soient mis au service de cette

3. ASPECTS DE LA QUALITÉ

La figure 1 est tirée de Lee et Allen (2001). Elle résume bien notamment (côté gauche) trois cadres en place d'évaluation de la qualité. On n'utilise pas partout les mêmes descripteurs, mais fonctionnellement le message est le même : il y a bien plus que l'exactitude à la qualité. C'est une vue qui est aujourd'hui largement acceptée, mais il n'y a pas si longtemps, l'examen de la qualité des statistiques visait plus particulièrement leur exactitude et la variabilité d'échantillonnage.

Il existe bien des façons de produire des statistiques officielles : d'après des données modélisées ou des produits analytiques basés sur les données de recensements et d'enquêtes par sondage. En Australie, nous faisons un plus grand usage qu'il y a cinq ans des données administratives, des systèmes de comptabilité (liés aux comptes nationaux), ainsi que de modèles et d'autres méthodes analytiques pour réaliser nos produits statistiques. Les défis en matière de qualité varient selon les moyens auxquels on recourt pour rassembler les données statistiques.

Les relations avec les autres organismes nationaux et les organismes statistiques se situent au cœur même des efforts d'amélioration de la statistique officielle à l'ABS. L'organisme adhère aux normes internationales et tire parti de la richesse des compétences qui s'incarnent dans ces normes. Il a par ailleurs l'obligation d'apporter une contribution à l'élaboration de normes. Ce faisant, il essaie de tenir compte des intérêts non seulement de l'Australie, mais aussi de toute la région Asie-Pacifique. Comme l'activité économique se mondialise sans cesse et que les visées sociales s'étendent à l'échelle du monde, la comparabilité des statistiques australiennes et de celles d'autres pays représente un important facteur de qualité. L'ABS entretient des relations étroites avec un grand nombre d'organismes étrangers. Il est heureux que les défis à relever soient souvent collectifs et qu'on ait de grands avantages à mettre les expériences en commun avec les autres organismes statistiques.

La formation statistique tient un grand rôle dans le maintien et l'amélioration de la qualité. L'ABS est constamment en quête de nouvelles méthodes plus efficaces de perfectionnement professionnel. Un élément de choix de son système de gestion du rendement est l'accent mis sur la constatation et la satisfaction des besoins individuels de perfectionnement.

L'appuyant sur des guides, des systèmes et des programmes de formation.

(i) Il existe bien des façons de produire des statistiques officielles : d'après des données modélisées ou des produits analytiques basés sur les données de recensements et d'enquêtes par sondage. En Australie, nous faisons un plus grand usage qu'il y a cinq ans des données administratives, des systèmes de comptabilité (liés aux comptes nationaux), ainsi que de modèles et d'autres méthodes analytiques pour réaliser nos produits statistiques. Les défis en matière de qualité varient selon les moyens auxquels on recourt pour rassembler les données statistiques.

Il y a plusieurs messages du côté droit de la figure 1.

statistiques. Les grandes statistiques macro-économiques doivent recevoir le feu vert des responsables de la comptabilité nationale à des réunions spéciales d'auto-risation de diffusion. Ces responsables contractent de ce fait l'obligation d'utiliser telles que les données produites dans l'élaboration de leurs comptes, d'où une plus grande cohérence entre les comptes nationaux et les sources des données. Plus généralement, les artisans de la comptabilité nationale procèdent à une confrontation des sources de données au moyen d'un système d'entrées-sources pour calculer les estimations de comptabilité nationale. Ce nouveau cadre méthodologique a permis de produire des comptes nationaux plus cohérents. En outre, l'exercice de comparaison et de rapprochement des données détaillées a facilité le dépistage des lacunes statistiques. L'information sur la qualité est communiquée en retour aux équipes de collecte des données économiques avec pour résultat une meilleure concentration des efforts de relèvement de la qualité des données de base.

Une importante initiative cultivée par l'ABS en matière d'amélioration de la qualité a été de créer un entrepôt d'information pour la gestion et le stockage de toutes ses données publiables. En réunissant tous les ensembles de données en un fonds d'information unique, cet entrepôt permet aux statisticiens de l'organisme de contrôler les statistiques issues des diverses activités de collecte. De plus, les publications de toutes sortes produites sur papier ou sur support électronique le sont à partir d'un même fonds d'information, le but étant de garantir la cohérence lorsque les mêmes données sont diffusées dans des produits différents ou à des moments différents.

Un autre élément important de gestion de la qualité est la documentation. Une bonne documentation appuie les exercices d'examen et facilite la diffusion de renseignements sur la qualité parmi les utilisateurs, lesquels sont ainsi en mesure de juger de l'adéquation des données aux usages qu'ils envisagent. Dans le cadre de l'initiative de l'entrepôt d'information, l'ABS peut aujourd'hui faire appliquer des normes de documentation des métadonnées où sont décrits les concepts, les définitions, les classifications et la qualité.

Un service statistique utile et sérieux ne doit pas se contenter de fournir des données à sa clientèle. L'ABS a renforcé récemment ses capacités d'analyse. Il a chargé une équipe d'analystes d'élaborer de nouvelles mesures des concepts socio-économiques, d'examiner les liens entre les variables et d'établir le prototype de nouveaux produits analytiques. On s'attend à ce que ce programme élargi d'analyse soit riche en retombées sous forme de renseignements sur les données manquantes et les problèmes de qualité.

2.5 Examen et évaluation des activités statistiques

Chaque secteur de l'ABS a des tâches d'étude et d'amélioration continues de la qualité. Dans les secteurs de

Ces dernières années, l'organisme a fait de grands progrès grâce à l'application de bonnes pratiques uniformes à toutes ses enquêtes. Ainsi, dans les enquêtes-entreprises qui reposent sur le registre des entreprises, le tirage des bases de sondage se fait à la même date chaque trimestre. On se sert d'une méthode d'estimation commune pour garantir une collecte cohérente et exhaustive de données dans tous les cas. On dispose de règles uniformes pour la mise à jour des bases de sondage et les travaux de collecte et d'estimation; des systèmes généralisés de traitement viennent soutenir l'application de ces règles. On a aussi des méthodes normalisées pour tenir compte des « nouvelles entreprises » non encore incluses dans les bases de sondage. L'ABS est donc en mesure d'accroître la cohérence des estimations produites d'après diverses enquêtes auprès des entreprises.

Dans le cas des enquêtes-ménages, on exploite un système d'échantillons principaux depuis le milieu de la décennie 1960. Ce système est régulièrement mis à jour après les recensements quinquennaux. Il a été à la base même des garanties d'exactitude des statistiques tirées des enquêtes auprès des ménages.

La qualité est chose plus facile dans les enquêtes où des systèmes informatiques appuient les bonnes pratiques adoptées. L'ABS a investi dans des systèmes généralisés. Il en a élaboré pour toutes les grandes étapes de traitement des données tant des enquêtes-entreprises que des enquêtes-ménages, qu'il s'agisse de gestion de bases de sondage, de saisie et de vérification de données, d'imputation, d'estimation ou d'agrégation.

L'ABS s'en tient à une démarche rigoureuse d'amélioration continue de la qualité partout où il y a lieu de le faire. Le recensement australien de la population est un exemple classique de relèvement de la qualité, car on y applique une stratégie de mesure de cette qualité et associe tout le personnel à l'étude et à la solution des problèmes de qualité. C'est une démarche qui s'est révélée très féconde au centre de traitement de données pour le recensement de 1996 et, dans d'autres cas, ce centre a fait d'importants économistes budgétaires et des gains de qualité et de rapidité de traitement. Le principe d'une amélioration continue de la qualité préside aussi au codage qui se fait au registre des entreprises et à bien d'autres activités de l'ABS.

Lorsque les activités de collecte débouchent sur des données d'autres données de l'organisme et à des données produits, chaque groupe spécialisé se doit de contrôler ses données pour une vérification de cohérence de ses

recours très fréquents à des outils de tests cognitifs à l'ABS et la création d'un laboratoire d'essai de questionnaires ont aidé à améliorer la qualité et à alléger le fardeau de réponse. Des normes d'évaluation et d'évaluation de formulaires sont énoncées dans des manuels. Des spécialistes en conception de questionnaires assurent la promotion et le soutien de leur application.

L'ABS se sert de plans de sondage efficaces afin de réduire au minimum les tailles d'échantillon nécessaires pour atteindre un niveau précis d'exactitude et, donc, le fardeau global de déclaration. Il contrôle aussi la sélection qui se fait dans les diverses collections de données pour que la répartition de la charge de réponse soit plus équitable. Soucieux de tirer parti des réformes en cours du régime fiscal australien, il cherche toutes les occasions de rendre ses plans de sondage plus efficaces grâce à l'utilisation de données fiscales comme valeurs de référence et pour remplacer certaines données de collecte directe. Nous avons modifié la structure d'unités d'entreprise utilisée dans nos enquêtes pour qu'elle concorde avec celle utilisée pour la déclaration de revenus.

Pour les enquêtes-ménages, l'instauration d'un programme d'interviews assistés par ordinateur a permis de rationaliser les techniques d'interview, d'alléger le fardeau de réponse et d'améliorer la qualité des données recueillies.

2.4 Procédés qui donnent des produits de grande qualité

La qualité des statistiques de l'ABS est garantie par l'application de bonnes méthodes statistiques à tous les stades de la collecte de données et, entre autres, à l'étape de la conception. L'organisme (environ 120 personnes) qui relève directement du statisticien en chef. Cette division est chargée de veiller à ce que des méthodes solides et défendables soient appliquées à toutes les opérations de collecte et de regroupement des données. Un comité de consultation méthodologique, formé de spécialistes des milieux universitaires, procède à un examen indépendant des méthodes statistiques de l'organisme.

L'ABS s'emploie à élaborer des normes statistiques comprenant des classifications, des définitions d'éléments d'information, des classifications et des modules de questions. Toutes les enquêtes menées doivent respecter ces normes dont l'application est soutenue par des systèmes de gestion de données qui en facilitent l'accès et l'utilisation (par opposition à l'application de méthodes non normalisées).

Les méthodes d'échantillonnage et d'estimation relèvent de la division de la méthodologie. Dans la mesure du possible, on applique le principe du « plan global d'enquête », où on fixe les exigences d'exactitude en fonction de l'usage prévu des données et mesure cette exactitude par les erreurs attribuables ou non à l'échantillonnage. Par exemple, dans les enquêtes-entreprises, un

Il peut aussi y avoir des conflits et des compromis entre les divers aspects de la qualité. L'ABS se positionne en « haut de gamme » de l'exactitude sur le marché de l'information afin de protéger la préférence « marque » ABS. Cependant, s'il faut, par exemple, répondre à une demande urgente de données dans un nouveau domaine, il arrive que l'on doive sacrifier certains aspects de la qualité pour pouvoir diffuser des statistiques pertinentes, dans les délais prévus. Il y a néanmoins une « barre » sous laquelle nous ne voulons pas glisser. Comme il est probable que les nouvelles statistiques ainsi produites soient destinées à éclairer un débat ou un exercice décisionnel important, l'ABS les accompagne de déclarations très claires concernant leur exactitude afin que les utilisateurs comprennent bien comment ils peuvent s'en servir. À l'occasion, le Bureau distingue ces nouvelles statistiques des autres produits en les qualifiant d'« expérimentales » ou en les diffusant sous forme de documents d'information ou de documents hors série. Nous considérons cette caractérisation comme un moyen très important d'assurer une interprétation sûre de nos statistiques.

2.3 Relations fructueuses avec les enquêtes

Un organisme statistique officiel doit entretenir de bonnes relations, tout spécialement un climat de confiance, avec les enquêtes, s'il veut que ceux-ci collaborent ou lui fournissent des données de grande qualité. Sur ce plan, l'ABS expose l'importance de ses données pour la politique et le débat publics et les décisions des entreprises, applique une politique d'essai systématique de tous les questionnaires préalablement à leur utilisation dans une enquête, obtient l'appui des grands intervenants, allège le plus possible le fardeau imposé aux enquêtes – surtout en exploitant les données administratives autant que possible – et veille à protéger la vie privée et à sauvegarder la confidentialité des renseignements.

L'ABS surveille et gère la charge qu'il fait supporter tant aux ménages qu'aux entreprises. Il a en outre créé en son sein un centre d'information répondant à l'intention de l'un et l'autre de ces groupes. Il a en outre créé en son sein un centre d'information statistique qui coordonne les enquêtes auprès des entreprises à l'échelle des organismes publics (ce qui comprend l'ABS), réduit les chevauchements et s'assure que des statistiques d'une qualité acceptable sont produites.

L'ABS soutient tous ses formulaires et méthodes de collecte à des essais pour s'assurer que les données recherchées puissent être obtenues à un coût raisonnable pour les enquêtes et que les meilleurs méthodes de collecte disponibles soient employées. Dans le cas des enquêtes-entreprises, il conçoit des modèles d'unités, des classifications et des éléments d'information qui reflètent aussi fidèlement que possible la façon dont fonctionnent les entreprises. Il s'aligne maintenant de près sur leurs déclarations de revenus, ce qui facilite l'intégration des données d'enquête aux données recueillies à des fins fiscales. Dans le cas des enquêtes auprès des ménages, le

Le respect de valeurs fondamentales n'est qu'un facteur du maintien d'une culture de la qualité. À la partie 2 sont exposées les principales mesures que prend l'ABS à cette fin.

On reconnaît largement aujourd'hui que la qualité est bien plus qu'une question d'exactitude des données (voir, par exemple, Brackstone (1999) et Carson (2000)). À la partie 3, les différents aspects de la qualité sont passés en revue avant d'indiquer, à la partie 4, quelques-uns des grands défis que devra relever le moyen terme l'Australian Bureau of Statistics (ABS) dans ce domaine. Nombre de ces défis se posent aussi à d'autres organismes statistiques nationaux.

2. VERS UN SERVICE STATISTIQUE DE GRANDE QUALITÉ

À l'ABS, l'assurance de la qualité est une tâche qui incombe à tout le personnel. On n'y trouve pas de groupe central chargé de la « gestion de la qualité », bien que la division de la méthodologie soit appelée à nous servir de conscience dans ce domaine, rôle qu'elle assume avec un enthousiasme parfois contrastant. Il s'agit pourtant d'un bon signe, car elle suscite ainsi des débats sur certaines questions de qualité épineuses. Il importe très nettement que la haute direction appuie un rôle de ce genre.

On peut ranger dans six grandes catégories les stratégies clés de maintien d'une grande qualité :

- haute crédibilité de l'ABS et de ses produits;
- maintien de la pertinence des produits de l'organisme;
- relations fécondes avec les enquêtes;
- procédés qui donnent des produits de grande qualité;
- examen et évaluation périodiques des activités statistiques;
- personnel compétent et animé d'une volonté d'assurer la qualité des produits de l'ABS.

2.1 Haute crédibilité

La crédibilité est primordiale pour une utilisation fructueuse des statistiques officielles. Elle naît d'un système statistique qui éclaire objectivement les gens sur l'état de l'économie et de la société d'un pays.

L'encadrement législatif de l'activité de l'ABS est un grand préalable de l'intégrité de la statistique officielle australienne. La loi confère au statisticien en chef (c'est-à-dire le premier dirigeant de l'ABS) une indépendance considérable qui aide à garantir que l'ABS soit impartial et à l'abri de toute ingérence politique, et qu'il soit perçu comme tel. Précisons que l'indépendance du statisticien en chef lui permet de demeurer objectif lorsqu'il établit le programme de travaux statistiques et détermine

Nos politiques consistant à faire part d'avance des dates de publication des grands indicateurs économiques, ce qui ne permet qu'une prédiffusion fort limitée des publications (dont les détails appartiennent au domaine public), et à mettre des services spéciaux d'information statistique à la disposition de tous renforcent l'objectivité de l'ABS, ainsi que la perception de celle-ci.

2.2 Maintien de la pertinence des produits de l'ABS

Bien sûr, il existe parfois un conflit entre le fait de répondre à l'évolution des besoins dans le domaine des politiques, d'une part, et celui d'assurer la continuité d'un système statistique permettant un contrôle objectif du rendement, d'autre part. La haute direction de l'ABS cultive les contacts personnels avec les principaux utilisateurs afin de se renseigner sur les questions de politiques et les nouveaux domaines d'intérêt économique, social ou écologique. On prévoit notamment des réunions régulières avec les représentants supérieurs des organismes publics responsables des politiques. Les directeurs de nos bureaux dans les États australiens en font de même avec les hauts représentants de ces derniers. Les renseignements ainsi recueillis facilitent la planification stratégique et l'examen des programmes statistiques nationaux.

L'ABS dispose de divers autres moyens de communiquer avec les utilisateurs de ses statistiques pour s'assurer que les produits sont adaptés à leurs besoins. Ainsi, les groupes consultatifs qui représentent les utilisateurs et les spécialistes des divers domaines offrent de précieux conseils quant à la façon d'orienter nos activités statistiques.

L'importance d'une culture de la qualité

DENNIS TREWIN¹

RÉSUMÉ

La réputation d'un organisme statistique national (OSN) dépend très largement de la qualité du service qu'il donne. La qualité doit être une valeur fondamentale bien ancrée dans la culture de l'organisme : offrir un service de grande qualité doit être chose naturelle. Le présent document évoque ce qu'on doit entendre par un service statistique de grande qualité. Il examine en outre les facteurs importants qui garantissent une culture de la qualité dans un OSN. Il y est brièvement question en particulier des activités et des expériences de l'Australian Bureau of Statistics sur ce plan.

MOTS CLÉS : Amélioration continue de la qualité; organisme statistique national; culture.

1. INTRODUCTION

Fellegi (1996) présente le solide argument selon lequel la confiance manifestée dans un organisme statistique national ne réside dans rien d'autre que la façon dont la plupart des utilisateurs jugent de la qualité de ses produits statistiques.

« La crédibilité joue un rôle fondamental dans la détermination de la valeur pour les utilisateurs de ce produit spécial que l'on appelle l'information statistique. En fait, peu d'utilisateurs peuvent directement valider les données que diffusent les organismes statistiques. Les usagers doivent s'en remettre à la réputation du fournisseur de données. Comme l'information à laquelle on ne croit pas est inutile, il s'ensuit que la valeur foncière et l'exploitabilité de l'information dépendent directement de la crédibilité du système statistique. Cette crédibilité peut être remise en cause en tout temps pour deux grands motifs, parce que les statistiques sont issues d'une méthodologie peu appropriée ou que l'organisme est soupçonné de partialité politique [TRANSDUCTION]. »

Il n'y aura pas de confiance sans une bonne culture. Le mot « culture » est porteur d'un grand nombre de significations, mais je l'interprète pour ma part comme la « façon de faire les choses ». Les valeurs fondamentales tiennent une grande place dans ce phénomène. Il ne peut tout simplement s'agir de maximes que l'on affiche au mur. Il faut une compréhension. Il faut que la culture s'incarne dans les comportements, plus particulièrement de ceux des dirigeants de l'organisme.

Confiance des fournisseurs de données – nous faisons avec les enquêtes un pacte par lequel nous leur demandons de nous fournir des données exactes, tandis que nous veillons pour notre part à la confidentialité des données qu'ils nous communiquent; nous réduisons le plus possible le fardeau et le dérangement imposés aux enquêtés. dans les limites d'exigences statistiques justifiées.

Professionnalisme – l'intégrité de nos statistiques tient à nos normes professionnelles et déontologiques; nous appliquons les normes professionnelles les plus élevées dans tous les aspects de l'activité statistique de l'ABS.

Accès général – nous produisons des statistiques au bénéfice de tous les Australiens et nous faisons en sorte que tous les utilisateurs jouissent d'un même accès aux données.

Intégrité – nos données, notre analyse et notre interprétation devraient toujours être objectives, et nous devrions publier des statistiques pour toutes nos collectivités de données; notre système statistique peut librement être examiné, car il repose sur des principes et des pratiques statistiques sains.

Pertinence – les contacts réguliers avec les gens qui peuvent agir sur les politiques et une bonne planification statistique – qui exige une bonne intelligence des besoins de statistiques, présents et futurs – sont essentiels, tout comme la nécessité de produire des statistiques actuelles et rattachables à d'autres statistiques.

L'Australian Bureau of Statistics (ABS) insiste beaucoup sur l'adhésion à ses valeurs fondamentales. Plus que toute autre chose, ces valeurs nous distinguent des autres organismes d'enquête en Australie. Les voici :

MOTS CLÉS : Amélioration continue de la qualité; organisme statistique national; culture.

- Techniques d'enquête, décembre 2002
- KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*. 14, 31-46.
- PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*. 35, 365-384.
- RAO, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*. 25, 2, 199-212.
- REDDERN, P. (1989). L'expérience européenne relative à l'utilisation des données administratives pour recenser la population : questions d'ordre politique. *Techniques d'enquête*. 15, 1, 85-103.
- SCHABELE, W.L. (1996). (Ed.) *Indirect Estimators in U.S. Federal Programs, Lecture Notes in Statistics*. New York: Springer-Verlag, 108.
- SINGH, M.P., GAMBINO, J. et MANTTEL, H.J. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*. 20, 1, 3-23.
- ZAYATZ, L., STEEL, P. et ROWLAND, S. (2000). Disclosure limitation for Census 2000. *Proceedings of the American Statistical Association, Section on Government Statistics and Section on Social Statistics*. 67-71.

individuelles suffisantes sont fiables de façon inhérente dans le cas des grandes régions (sans tenir compte du biais pour le moment), ne semble pas pouvoir être soutenue pour les régions plus petites. À moins de disposer d'un recensement ou d'une source administrative à jour, comportant une couverture complète, le BSN doit avoir recours à une forme de mesure axée sur un modèle pour produire des estimations. Étant donné que les divers modèles permettent de produire des estimations différentes, cela amène un degré d'arbitraire dans les estimations, et peut être perçu par certains comme nuisant à l'objectivité d'un BSN et à ses méthodes. Le principe fondamental d'ouverture et de transparence au sujet des méthodes, y compris le choix des modèles utilisés et les répétitions des différentes hypothèses, est encore plus important dans le domaine des estimations régionales.

Par-dessus tout, un BSN doit s'attendre à ce que les estimations régionales fassent l'objet d'une attention beaucoup plus étroite que nombre d'estimations pour les grandes régions. Même si les estimations pour les grandes régions reçoivent une attention plus marquée, peu de personnes ont la capacité de confirmer ou de réfuter une estimation au niveau national. Toutefois, au niveau local, nombreux sont ceux qui pensent savoir de quoi il retourne. Et de façon générale, les estimations régionales ne fonctionnent pas uniformément bien pour toutes les régions. Une méthode qui fonctionne bien en moyenne n'est pas à l'abri des critiques dans les régions où elle n'a pas bien fonctionné, à moins que cela n'ait été à l'avantage de ces régions! Le BSN doit être prêt à faire face au double obstacle que présentent des estimations plus faibles faisant l'objet d'une attention plus étroite.

Et comme si cela n'était pas suffisant, les considérations en matière de confidentialité sont plus grandes au niveau régional. Du fait même que des estimations sont produites pour des régions, il est possible que des personnes puissent être identifiées, même si le BSN a pris suffisamment de précautions pour prévenir une telle divulgation. Certains utilisateurs des données régionales dans le secteur du marketing continuent à accentuer le problème en soulignant dans leurs annonces qu'ils peuvent cibler le courtier qu'ils envoient aux ménages, selon les caractéristiques des personnes ou des ménages, lorsqu'ils utilisent les données régionales pour faire une distinction entre les divers quartiers. Certaines méthodes d'estimations régionales nécessitent le couplage d'enregistrements, ce qui soulève aussi des questions du point de vue de la protection des renseignements personnels. Encore une fois, une politique d'ouverture et un examen soigné de toutes les applications, à un niveau supérieur et avant même que ces applications soient mises en oeuvre, est nécessaire pour s'assurer que les avantages pour le public surpassent les intrusions dans la vie privée des gens.

En dépit de ces difficultés, la demande de données régionales demeure élevée, la technologie offre de nouvelles approches pour la gestion et la diffusion des

REMERCIEMENTS

données régionales, et les travaux méthodologiques relatifs aux estimations régionales constituent un secteur actif de recherche parmi les statisticiens. Même si la production des données régionales ne constitue généralement pas la première priorité d'un BSN, la pertinence des programmes statistiques sera grandement améliorée si les BSN peuvent répondre aux besoins les plus importants en matière de données régionales.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

ALEXANDER, C.H. (2002). Les échantillons successifs de Leslie. *Kish et l'American Community Survey. Techniques d'enquête*. 28, 1, 39-46.

BAYARRI, M.J., et BERGER, J.O. (2000). *P Values for composite null models. Journal of the American Statistical Association*. 95, 452, 1127-1142.

BRACKSTONE, G. (1987). Utilisation des dossiers administratifs à des fins statistiques. *Techniques d'enquête*. 13, 1, 35-51.

DURR, J.-M., et DUMAIS, J. (2002). La rénovation du recensement français. *Techniques d'enquête*. 28, 1, 47-53.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *Journal of American Statistical Association*. 74, 269-277.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on Statistical Disclosure Limitation Methodology (Statistical Policy Working Paper #22). Washington, D.C., Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

GAMBINO, J., et DICK, P. (2000). Small area estimation practice at Statistics Canada. *Statistics in Transition*. 4, 597-610.

GOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*. 9, 55-93.

ISNARD, M. (1999). *Alternatives to Traditional Census Taking: The French Experience*. Paris: INSEE.

JABINE, T.B. (1993). Statistical disclosure limitation practices of united states statistical agencies. *Journal of Official Statistics*. 9, 2, 427-454.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*. 16, 1, 67-86.

d'estimation peuvent devoir reposer sur des techniques d'analyse spatiale, comme l'établissement de cartes en courbes de niveau ou l'interpolation.

Les préoccupations en matière de protection des renseignements personnels et de confidentialité liées aux données environnementales dépendent de la source de ces données. Les données recueillies auprès des ménages ou des entreprises, même si elles comportent des mesures matérielles, sont protégées par les mêmes règles de confidentialité que les autres données provenant de ces sources. Des mesures directes du stock de ressources naturelles ou de la qualité de l'environnement ne soulèvent pas de telles préoccupations. La représentation cartographique des modèles spatiaux peut constituer une façon de surmonter certaines des frustrations analytiques découlant de la suppression de données régionales. Les cartes choroplèthes (qui illustrent la répartition des variables ou des caractéristiques grâce à des zones colorées ou ombrées correspondantes) peuvent représenter de façon explicite les fourchettes implicites dans les lignes et les colonnes des tableaux publiés.

Les flux transfrontaliers de polluants et leurs effets globaux donnent un caractère international aux données environnementales. La coopération de pays voisins est nécessaire pour faire en sorte que les frontières nationales n'empêchent pas l'analyse des répercussions des procédés matériels qui dépassent ces frontières.

En résumé, la dimension régionale est particulièrement importante dans le cas des données environnementales, non seulement parce que le caractère local représente fréquemment le centre d'intérêt, mais aussi parce que les données doivent souvent être réagrégées selon des régions géographiques qui se prêtent mieux à l'analyse environnementale, comme les écosystèmes et les bassins hydrographiques.

7. QUESTIONS D'ORGANISATION ET DE DIFFUSION

La plupart des BSN sont organisées en secteurs spécialisés. La production d'estimations régionales recoupe ces secteurs, et elle nécessite le soutien du personnel de la Géographie pour l'infrastructure géographique, de la Méthodologie pour les méthodes d'estimation et d'évaluation, et parfois de la Diffusion, pour l'intégration des données d'un secteur à l'autre. La question de l'organisation des estimations régionales à l'intérieur des BSN se pose donc.

Le fait de confier aux secteurs spécialisés la responsabilité de la gestion des estimations régionales les concernant, avec l'appui du personnel de la Méthodologie et de la Géographie, au besoin, représente un choix naturel, étant donné que ces secteurs sont probablement davantage au courant des exigences et des limites s'appliquant aux données dans leur domaine. Ce qui présente davantage un problème, c'est la façon de regrouper les données

régionales en vue de leur diffusion aux utilisateurs. Qui devrait être responsable d'obtenir les données des divers secteurs spécialisés pour une région particulière? Devrait-il s'agir d'un programme régulier, ou devrait-on parfois le faire « sur demande »? Divers modèles s'offrent, et Statistique Canada les a essayés, pour la plupart, au fil des ans.

À un moment donné, une division axée sur les statistiques régionales ou urbaines a été mise en place pour assurer le caractère régional des données statistiques. À un autre moment, le programme de recensement, qui constitue évidemment la source la plus importante de données régionales, a été à l'avant-plan de la production de profil de données régionales. Parfois aussi, on a eu recours à un projet faisant intervenir plusieurs divisions pour gérer un programme de profils relatifs aux districts électoraux ou à d'autres régions géographiques. Parallèlement, les employés des bureaux régionaux ont joué un rôle clé pour regrouper des données régionales en réponse à des demandes des clients. Aucune de ces façons de faire n'est idéale. La production de profils a toujours nécessité beaucoup de travail, ainsi qu'une vaste compréhension des domaines spécialisés et de nombreuses recherches et manipulations de données. En dépit de l'existence de régions géographiques normalisées, la combinaison de données fondées sur plusieurs bases géographiques différentes pose généralement un problème. Le contrôle de la qualité présente un défi majeur lorsqu'on veut s'assurer qu'un grand nombre de petites régions sont correctement

appartées et re-groupées.

Les profils sur papier planifiés au préalable n'ont jamais connu un très grand succès, ce qui fait que l'on a privilégié une stratégie répondant au maximum aux demandes des clients, au fur et à mesure qu'elles se présentent. Du fait des progrès récents de la technologie et de la couverture plus large des données régionales dans la base de données du Bureau, une approche d'avantage automatisée est possible. Une composante du site Internet de Statistique Canada (www.statcan.ca), appelée Profil statistique des communautés canadiennes et fondée pour une large part sur les données du recensement, constitue notre tentative la plus récente pour rendre les données régionales plus accessibles, et semble présager des orientations à venir dans ce domaine. Certaines données sur la santé pour des régions socio-sanitaires ont déjà été incorporées et va contredire l'ajout d'autres données ne provenant pas du recensement.

8. CONCLUSIONS

La production de statistiques régionales par un BSN soulève des problèmes qui sont différents au point de vue qualitatif de ceux découlant de la production régulière de données nationales et provinciales ou de données concernant d'autres grandes régions. La théorie statistique selon laquelle les données fondées sur des mesures

différence lorsque l'on s'intéresse aux totaux provinciaux seulement. En général, toutefois, les règles de répartition géographique doivent être déterminées avant d'envisager la production d'estimations régionales pour les activités des entreprises, et dans le cas de certains aspects de ces activités, les estimations régionales n'ont pas de sens au niveau conceptuel.

Même si pour les enquêtes auprès des ménages, le principal obstacle quant à la production d'estimations régionales est la taille de l'échantillon, pour les enquêtes auprès des entreprises, ce sont les considérations de confidentialité qui constituent généralement l'obstacle majeur. Plus la région est petite, plus il y a de chances qu'une ou quelques entreprises majeures dominent un secteur d'activité particulier, ce qui empêche la production d'estimations pour cette région en raison des risques de divulgation. Les méthodes de vérification des produits statistiques relatifs aux entreprises qui permettent de déterminer les risques possibles du point de vue de la divulgation sont assez bien développées (FCSM 1994), mais elles nécessitent une attention constante de la part des BSN. Le problème de la confidentialité est moins important pour les branches d'activité caractérisées par des petites unités, et il peut s'agir des mêmes branches d'activité pour lesquelles les problèmes conceptuels abordés au paragraphe précédent ne sont pas aussi graves. Pour ces branches d'activité, les considérations relatives à la taille de l'échantillon peuvent en fait constituer le facteur restrictif, auquel cas on peut avoir recourus aux familles de méthodes décrites dans la section précédente.

Il existe un troisième domaine de divergence par rapport aux données sur les personnes, à tout le moins pour les pays qui ne tiennent pas de registre de la population, à savoir l'existence de nomenclatures relativement à jour des entreprises. Cela fournit non seulement une base d'échantillonnage et une source de données auxiliaires pour l'estimation, mais aussi une source possible d'estimations directes de la démographie des entreprises, à tout le moins sur une base annuelle. Dans nombre de pays, le registre des entreprises est mis à jour grâce aux opérations transmises par le système d'imposition des entreprises, lequel constitue en soi une source annuelle de données administratives s'appuyant sur un recensement sur les activités des entreprises. Toutefois, l'utilisation des données fiscales nécessite encore un examen soigneux des questions conceptuelles et géographiques ainsi que des questions de confidentialité soulevées ci-dessus.

6. STATISTIQUES ENVIRONNEMENTALES

L'industrie est axée sur les ressources. Certaines données environnementales sont recueillies auprès des ménages ou des particuliers (par exemple, sur les pratiques de recyclage, la consommation de carburant) et leurs possibilités en tant que source de données régionales sont sujettes à des considérations déjà soulevées à la section 4. D'autres données environnementales (par exemple, sur la production de déchets, les dépenses pour la protection de l'environnement, l'utilisation des ressources naturelles) proviennent des entreprises et sont associées aux considérations énoncées à la section 5. Toutefois, une part importante des données environnementales sont obtenues à partir d'enquêtes sur le terrain (par exemple, géologiques, physiographiques et hydrographiques), de mesures au moyen d'instruments (par exemple, température, qualité de l'air, qualité de l'eau, épaisseurs de la couche d'ozone), et d'observations directes (par exemple, utilisation du territoire). Des considérations différentes régissent les rapports qui existent entre ces sources de données et les données régionales.

Du fait que les données environnementales ne respectent pas les limites administratives, la nécessité d'une infrastructure géographique souple, qui est soulignée à la section 3, révèle une importance particulière dans ce cas. La détermination géographique des limites régionales est nécessaire pour regrouper les données selon des unités géographiques qui se prêtent davantage à l'analyse environnementale. Par exemple, la production de déchets urbaine à un certain type d'activité agricole peut être agrégée pour l'ensemble des producteurs avoisinants un bassin fluvial. Les unités géographiques environnementales sont, soit définies au préalable (écozones, bassins de drainage), soit dictées par des événements spéciaux (des régions couvertes de différentes épaisseurs de glace, terres inondées par des pluies abondantes ou par la fonte des neiges au printemps). Dans certains cas, la région étudiée peut se limiter à un emplacement très petit, par exemple, un parc.

Des données sur la quantité ou la qualité matérielle peuvent être difficiles à agréger ou à résumer. Dans certains cas, des données de source ponctuelle, comme les mesures de la qualité de l'air, ne peuvent être considérées comme représentatives d'unités géographiques plus importantes. La qualité de l'eau peut faire l'objet de résumés ou de comparaisons, grâce à un indicateur, par exemple, le nombre de jours d'ouverture des plages pour la baignade, mais ne peut pas constituer simplement une agrégation ou une moyenne des données sur la qualité. Pour nombre de mesures, l'accent est mis sur les changements qui se produisent au fil du temps, plutôt que sur les comparaisons régionales. Dans d'autres cas, des techniques d'échantillonnage et

repères en fonction des nouvelles estimations courantes. Essentiellement, toutes ces méthodes nécessitent qu'un équilibre soit établi entre ces trois types d'estimations : a) variance élevée, mais estimations d'enquête directe courante non biaisées pour la région en question; b) faible variance pour les estimations d'enquête courante à l'égard d'une région voisine plus grande comprenant la région en question; c) estimations de type recensement pour la même région à partir des données administratives récentes, ou d'un recensement passé, qui peuvent contenir un biais indétectable, en raison de la source et du délai écoulé. Toutes les données auxiliaires disponibles peuvent être intégrées pour améliorer la précision de chaque estimation de composante. La façon de combiner ces trois types d'estimations est déterminée par le choix et les paramètres du modèle.

En résumé, les méthodes énoncées dans la présente section et dans la section précédente ont essentiellement pour effet de réduire la variance, du fait qu'elles utilisent davantage de données, mais elles peuvent entraîner un biais en raison de l'utilisation d'hypothèses modèles qui ne seront jamais exactement correctes. Il est très important d'analyser le rendement de ces méthodes avant de les utiliser, par exemple, en procédant au processus d'estimation au cours d'une année de recensement, lorsque des estimations directes sont disponibles aux fins de la comparaison, et de façon périodique par la suite. La vérification des modèles est un domaine de recherche en développement (Bayarri et Berger 2000). Pour des descriptions plus détaillées des méthodes disponibles dans cette catégorie, voir, par exemple, Purcell et Kish (1979); Fay et Herriot (1979); Ghosh et Rao (1994); Singh et coll. (1994); Schabale (1996); Rao (1999) et Gambino et Dick (2000).

4.4 Les recensements continus

Une solution de remplacement innovatrice au recensement est envisagée dans au moins deux pays. La méthode de production de données régionales à partir d'un échantillon cumulé important est préconisée depuis longtemps par Leslie Kish, comme solution de remplacement au recensement traditionnel (Kish 1990, 1998). On procède au cumul des résultats de l'enquête sur échantillon, c'est-à-dire qu'au cours d'une période prolongée (par exemple, une décennie), chacune des régions plus petites pour laquelle des estimations sont nécessaires est incluse une fois dans l'échantillon, ce qui permet la production d'une estimation directe pour cette région au moins une fois pour chaque période. Des régions de plus en plus grande (agrégation des régions plus petites) sont représentées plus souvent dans l'échantillon, ce qui permet des estimations plus fiables ou plus fréquentes pour ces régions. Dans le cas des régions encore plus grandes, par exemple, des provinces ou l'ensemble du pays, l'échantillon cumulé est suffisant pour fournir des estimations annuelles fiables, ou plus fréquentes, selon certains niveaux de détail. L'approche peut être envisagée, parallèlement à des recensements

Le recensement continu permet d'éviter l'élaboration de modèles, mais il repose sur le principe que des estimations non biaisées de moyennes sur plusieurs années, ou des estimations asynchrones pour diverses régions du pays, constituent des options de échange satisfaisantes aux estimations simultanées à un point donné dans le temps des recensements traditionnels. Le coût relatif constitue aussi un facteur clef, particulièrement dans les cas où un recensement de base est aussi effectué. Par ailleurs, cette approche tient compte du fait que les estimations de recensement peuvent dater de 12 ans lorsque les suivantes sont produites, à partir d'estimations annuelles fiables pour de nombreuses régions plus grandes, le contenu de ces estimations s'appliquant à celui des recensements du point de vue du niveau de détail. Elle fait aussi suite aux préoccupations croissantes concernant les difficultés et les coûts de plus en plus grands liés à la tenue d'un recensement traditionnel.

Cette approche fait l'objet d'un essai aux États-Unis, sous le nom d'American Community Survey (Alexander 1999), et en France, sous le nom de Recensement continu de la population (Isnard 1999).

5. STATISTIQUES SUR LES ENTREPRISES

Les problèmes de production de données régionales pour les entreprises diffèrent à de nombreux égards de ceux relatifs aux données sur les personnes ou les ménages. Même si l'association d'une personne avec un « lieu

habitué de résidence » représente, pour la grande majorité de la population, un concept (concept qui est peut-être moins bien défini avec l'augmentation du nombre de résidents secondaires, de l'incidence des absences prolongées pour des destinations plus au sud, et des conditions de logement plus souples), relativement clair et non ambigu, la question de la répartition géographique des diverses caractéristiques des entreprises est moins évidente dans nombre de cas. Pour ce qui est des entreprises qui ne comptent qu'un établissement, et dont toutes les activités se déroulent à un seul emplacement, il n'y a pas de problème conceptuel, même s'il peut y avoir un problème pratique lorsque la source des données est un fichier administratif qui comporte, par exemple, l'adresse d'un comptable, plutôt que l'adresse de l'entreprise. Dans le cas de certaines variables, pour celles du secteur du transport, ou pour certaines industries de service). Toutefois, dans le cas des variables comme le revenu et les bénéfices, des questions réelles peuvent se poser quant à leur répartition géographique pour les entreprises comptant plusieurs établissements. Plus la région géographique est grande, moins le problème est grave, l'emplacement dans une province ne faisant pas de

En ce qui a trait aux données administratives en général, le statisticien doit tirer parti de ce qui est disponible (même s'il peut influer sur le contenu à plus long terme), tenir compte des écarts entre les concepts, la définition ou la couverture des fichiers administratifs et les objectifs statistiques, et évaluer les problèmes de déclaration ou de codage des enregistrements. Sous réserve de ces précautions, les données administratives peuvent constituer une source géographique valable de données régionales (Brackstone 1987).

4.2 Données des enquêtes par échantillon

Le problème que présentent les données des enquêtes par échantillon comme sources de statistiques régionales a trait à la taille de l'échantillon. Il arrive fréquemment que le nombre de cas compris dans l'échantillon ne soit pas suffisant dans une région pour permettre des estimations directes fiables. Parfois même, il est impossible de produire des estimations. Dans le cadre des enquêtes par échantillon nationales importantes, on peut concevoir des stratégies d'échantillonnage permettant d'assurer un niveau d'acceptation de précision pour des régions définies, par exemple, les régions infra-provinciales, sans diminuer de façon significative la fiabilité des estimations à des niveaux plus élevés (Singh, Gambino et Mantel 1994). Toutefois, il n'est pas possible de produire des estimations fiables pour des régions plus petites, ou pour des régions de taille similaire qui n'ont pas été prises en compte au moment du plan d'échantillonnage. Les échantillons plus grands sont utiles et peuvent permettre une estimation directe pour certaines des grandes régions, mais les budgets limitent généralement cette approche en tant que solution générale. Si aucune autre source de données n'est disponible, les statisticiens doivent se fier uniquement aux méthodes fondées sur des modèles, lesquelles nécessitent que l'on fasse des hypothèses au sujet des liens qui existent entre les données régionales et les autres données. Ces méthodes reposent souvent sur le transfert d'information, c'est-à-dire qu'elles consistent à emprunter de l'information d'autres secteurs de l'enquête sur échantillon pour augmenter le nombre d'unités qui contribuent aux estimations pour une région donnée. Le transfert peut se faire à partir d'autres périodes, d'unités d'échantillonnage extérieures à la région donnée, ou d'autres variables mesurées à partir de la même unité d'échantillonnage. Voici des exemples de ces cas. La plupart d'entre eux permettent d'élargir la gamme des estimations régionales qui peuvent être produites à partir des enquêtes comportant des échantillons relativement grands. Ils ne permettent toutefois pas de convertir de façon magique les enquêtes reposant sur de petits échantillons en sources importantes de données régionales.

1. Dans le cadre d'une enquête mensuelle, il est possible de combiner des données pour une région, pendant plusieurs mois consécutifs, en vue de produire des estimations directes d'une moyenne

mobile sur plusieurs mois pour la région. Cela peut par exemple permettre la production d'estimations trimestrielles, alors que des estimations mensuelles ne sont pas possibles.

2. On peut consentir à faire l'hypothèse que les moyennes ou proportions estimées pour une région plus grande s'appliquent également à une région plus petite qu'elle englobe. Si la taille de la région est connue, on peut obtenir une estimation en multipliant par la moyenne ou la proportion supposée. Cette hypothèse est plus réaliste pour des sous-groupes de population (par exemple, des groupes d'âge), plutôt que pour la population dans son ensemble. Dans ce cas, si la taille de chaque sous-groupe est connue pour la région, un estimateur synthétique peut être établi grâce à la multiplication des tailles et des moyennes supposées et à l'agrégation des résultats.

3. Si des variables connexes additionnelles sont disponibles à partir de l'enquête, des modèles plus élaborés peuvent être établis et comporter un lien entre la variable faisant l'objet de l'estimation et ces variables auxiliaires. Les paramètres du modèle peuvent être estimés à un niveau géographique plus élevé, lorsque l'échantillon est suffisant pour procéder à cette estimation de façon fiable. Le modèle est par la suite appliqué avec les paramètres estimés aux données pour la région déterminée.

Pour toutes ces approches, on manque de données de base fiables pour chaque région. Si de telles données étaient disponibles, par exemple, à partir d'un recensement récent ou de dossiers administratifs, elles pourraient être utilisées en combinaison avec l'une ou l'autre des sources pour produire des estimations plus fiables que chacune des sources isolément.

4.3 Sources combinées

Les méthodes qui comment des données de recensement ou des données administratives récentes et des données d'enquête sur échantillon courantes profitent du transfert de données extérieures à l'enquête. Elles nécessitent aussi des hypothèses sur les modèles, mais ces dernières sont souvent plus faibles (étant donné qu'elles reposent sur des hypothèses quant au changement par rapport au repère, plutôt qu'au sujet de niveaux absolus pour chaque région) et sont donc davantage acceptables, ou plus plausibles, que les données d'enquêtes sur échantillon prises isolément.

Une gamme variée de méthodes d'estimation (que nous ne décrivons pas ici) a été élaborée pour traiter cette situation. Certaines de ces méthodes peuvent être perçues comme une estimation des changements par rapport au repère le plus récent, d'autres comme une répartition d'estimations d'enquête sur échantillon à jour et fiable entre les composantes, à partir des données repères, et d'autres encore, comme une recalibration des anciens chiffres

directe des résultats de chaque dénombrement périodique. Au cours des périodes intercensitaires, les données de recensement peuvent être utilisées comme repères, comme base d'échantillonnage ou comme données auxiliaires, à l'égard d'autres sources de données qui sont disponibles. Ces utilisations sont décrites de façon plus détaillée à la section 4. Une solution de rechange innovatrice au recensement traditionnel est décrite à la section 4.4.

3. INFRASTRUCTURE GÉOGRAPHIQUE

Pour qu'un recensement national permette la production de données régionales précises, il faut disposer au préalable d'une infrastructure géographique des limites et d'une fonction de cartographie pour l'ensemble du pays. Une telle infrastructure nécessite que chaque logement soit lié à un emplacement géographique précis sur le terrain, le degré de précision étant déterminant du niveau de définition des régions. Même si la technologie moderne de positionnement global fait en sorte qu'il est possible de repérer chaque logement selon une paire précise de coordonnées, il suffit généralement, à des fins statistiques, de lier chaque logement d'une région urbaine à un côté d'îlot (c'est-à-dire un côté d'une rue entre deux intersections), ou à un immeuble, dans le cas des immeubles en hauteur. Dans les régions rurales, le niveau de précision choisi dépend des limites administratives et naturelles locales, l'utilisation de coordonnées précises pour chaque logement permettant d'assurer le maximum de souplesse.

L'infrastructure géographique, qui est nécessaire dans le cadre du recensement, est aussi utile pour la production de statistiques régionales à partir d'autres sources. Essentiellement, chaque point de données, quelle qu'en soit la source, doit être lié à un emplacement géographique, selon un niveau de détail suffisant pour permettre l'agrégation en régions présentant un intérêt statistique. Par exemple, si la source des données est un registre administratif, ou un registre des entreprises, l'adresse de chaque enregistré doit pouvoir être transposée en une paire de coordonnées géographiques, ou à tout le moins correspondre à la région dans laquelle se trouve cette adresse. Étant donné que les registres administratifs utilisent souvent les adresses postales, un fichier qui convertit les codes postaux en emplacements géographiques constitue un outil valable pour l'élaboration de données régionales.

Une infrastructure géographique précise et à jour, qu'elle soit établie par le BSN ou obtenue au niveau externe, est essentielle pour que l'on puisse choisir les régions pour lesquelles des statistiques seront produites, dans le cadre d'un programme de statistiques régionales.

Nous aborderons maintenant la question de la production de données régionales pour les personnes ou les ménages, au cours des périodes qui séparent les recensements. Il ressort clairement que l'existence d'un registre à jour de la population fait toute la différence en ce qui a trait aux possibilités et à la façon de faire. Nous nous limiterons aux cas où il n'existe pas de registre de la population régulièrement mis à jour.

Dans de tels cas, trois catégories principales d'approches s'offrent. La première consiste à utiliser des fichiers systèmes administratifs et qui visent à couvrir l'ensemble s'apparentant à ceux du recensement qui fournissent de façon précise les données des enquêtes sur échantillon et, grâce à exploiter les données des enquêtes sur échantillon et, grâce à des hypothèses modèles additionnelles, à produire des estimations pour des régions plus petites (quoique pas encore très petites) que celles visées par l'estimation directe découlant d'enquêtes. La troisième catégorie représente une combinaison de ces deux approches et utilise les données du recensement le plus récent. Dans les paragraphes qui suivent, nous passerons en revue certaines des caractéristiques de ces approches.

4.1 Fichiers administratifs

Parmi les fichiers administratifs qui comportent des possibilités statistiques au niveau régional figure le fichier annuel des déclarations de revenu des particuliers. Parmi les autres exemples, pour des populations plus restreintes, figurent les fichiers des titulaires de permis de conduire, des prestations de l'assurance-emploi ou des bénéficiaires de l'assurance-maladie. Dans les cas des données sur les revenus des particuliers, si chaque enregistré est lié à une région géographique ou à une région, des données peuvent être obtenues directement pour les régions, sous réserve des exigences relatives à la confidentialité (comme c'est le cas pour les données du recensement). Les caractéristiques disponibles ne dépassent généralement pas les variables démographiques et les variables de revenu, et la couverture se limite aux déclarants. Néanmoins, un tel fichier représente une source importante de données annuelles pour des régions relativement petites. La couverture de la population peut être améliorée grâce à l'imputation des personnes à charge pour lesquelles des déductions sont demandées dans la déclaration de revenu. Au Canada, la couverture de ces fichiers imputés s'apparente à celle du recensement, étant donné que la couverture des personnes à faible revenu qui doivent produire des déclarations de revenu pour obtenir des prestations d'aide sociale augmente.

4. STATISTIQUES RÉGIONALES SUR LES PERSONNES ET LES MÉNAGES - PÉRIODES INTERCENSITAIRES

couverture.

Dans le présent document, nous passons d'abord en

Dans la plupart des pays, le recensement de la popula-

d'estimations pour de très petites régions, des règles visant à empêcher la divulgation directe ou par recoupement des

Stratégies et approches aux statistiques régionales

GORDON J. BRACKSTONE¹

RÉSUMÉ

Les bureaux statistiques nationaux sont souvent appelés à produire des statistiques pour de petites régions géographiques, en plus d'assumer leur responsabilité principale, qui est de mesurer l'état du pays dans son ensemble et celui de ses grandes subdivisions. Cette tâche présente des défis différents de ceux qui existent dans le cadre de programmes statistiques visant principalement l'obtention de données nationales ou provinciales. Dans le présent document, nous étudions ces défis et déterminons des stratégies et des approches en vue de l'élaboration de programmes de statistiques régionales. La base importante que constitue le recensement de la population de même que le rôle de premier plan que joue une infrastructure géographique cohérente y sont mis en évidence. Des sources et des méthodes possibles en vue de la production de données régionales dans les domaines social, économique et environnemental y sont examinées. Des questions d'organisation et de diffusion y sont également abordées.

1. INTRODUCTION

La plupart des bureaux statistiques nationaux (BSN) ont essentiellement pour mandat de suivre les conditions sociales, économiques et environnementales au niveau national, ainsi qu'à l'échelon des principales unités administratives (provinces, États, régions métropolitaines importantes) à l'intérieur du pays. Toutefois, la demande de données à des niveaux géographiques plus restreints est toujours présente, particulièrement chez les administrations locales et les entreprises qui doivent prendre des décisions en matière d'investissement, de marketing et d'implantation, décisions qui nécessitent une connaissance des régions locales. Nous utiliserons le terme « statistiques régionales » pour parler des statistiques s'appliquant à des régions plus restreintes qu'un État, une province ou une région métropolitaine importante, soit toute une gamme de régions qui va des villes importantes aux villages ruraux, en passant par les quartiers urbains. Dans certains milieux on fait référence au concept plus large de « petits domaines »; cependant ici on parle strictement de petites régions géogra-

La portée des responsabilités des BSN en ce qui a trait aux statistiques régionales dépend de la répartition des responsabilités gouvernementales au sein d'un pays. Par exemple, dans certains pays, les administrations locales sont créées par les provinces et la responsabilité relative à leurs besoins statistiques sont du ressort des administrations provinciales. Toutefois, dans de nombreux pays, quelle que soit la répartition officielle des pouvoirs, on s'attend dans les faits à ce que le BSN répondre aux besoins de statistiques régionales, à partir de ses propres ressources, ou en collaboration avec d'autres niveaux de gouvernement. Le BSN doit, à tout le moins, établir les normes et le cadre s'appliquant aux données régionales, afin que celles-ci ne

Il existe quatre sources possibles de données statistiques régionales par les organismes statistiques. Les recensements régionaux par les organismes statistiques de la population consistent en une source traditionnelle. Les dossiers administratifs, y compris les registres nationaux qui couvrent la totalité ou la presque totalité d'une population définie, sont à de nombreux égards équivalents à un recensement. Les enquêtes nationales sur échantillon sont rarement de taille suffisante pour produire des données régionales directes-mais elles constituent une source valable d'information à jour qui peut être utilisée, sur la base de certaines hypothèses et en combinaison avec d'autres sources, pour produire des données régionales. Enfin, des études locales axées sur des régions particulières permettront de produire des données régionales, mais elles ne peuvent s'appliquer

et non comparables au niveau national. Du fait des budgets limités, le BSN est aux prises avec un compromis difficile entre des investissements à l'égard de statistiques nationales et la production de données régionales détaillées. Doit-il choisir de couvrir davantage de domaines ou de couvrir les domaines existants de façon plus détaillée, aux niveaux national et provincial, ou encore de fournir davantage de données régionales détaillées pour des sujets qui sont déjà couverts au niveau national. Il n'existe pas de formule établie pour résoudre ce problème. L'équilibre obtenu dépend pour une large part des besoins nationaux, des pouvoirs relatifs et de la tradition, et peut-être de certaines considérations statistiques en parallèle. Néanmoins, il existe une série de mesures et d'approches qu'un BSN doit envisager pour répondre dans la plus large mesure possible aux demandes de statistiques régionales, à partir d'un budget limité.

¹ Gordon J. Brackstone, Secrétaire Informatique, Statistique Canada, Ottawa, (Ontario), K1A 0T6, courrier électronique: brackgor@statcan.ca

Sirken considère l'estimation du volume de transactions qui ont lieu entre une population d'établissements et une population de ménages. Il compare une méthode fondée sur l'échantillonnage indirect des établissements par le biais des ménages qui font des transactions avec ces établissements à une méthode plus typique fondée sur l'échantillonnage des établissements avec probabilité proportionnelle à la taille. Il établit les estimateurs et les expressions du calcul de leur variance pour les deux méthodes, et les compare afin de préciser les situations où une méthode est meilleure que l'autre.

Rivest étudie le problème de la définition des limites de strates. L'algorithme de Lavalée-Hidroglou utilise habituellement supposé qu'on connaît les valeurs de la variable étudiée et qu'on les utilise pour déterminer les limites optimales des strates. Dans le présent article, Rivest relâche cette hypothèse et modifie l'algorithme de Lavalée-Hidroglou pour tenir compte d'une différence entre la variable de stratification et la variable étudiée grâce à l'utilisation de modèles qui établissent un lien entre ces deux variables. Puis, il intègre ces modèles dans l'algorithme de Lavalée-Hidroglou.

Dans leur article, Lu et Sitter décrivent le problème qui se pose lorsque la taille souhaitée de l'échantillon est inférieure ou à peine supérieure au nombre total de strates. Dans ces conditions, les méthodes habituelles de répartition de l'échantillon entre les strates ne sont pas toujours applicables. Une solution consiste à utiliser une méthode de programmation linéaire pour minimiser le manque prévu de « désirabilité » des échantillons à condition que soit respectée une contrainte de répartition proportionnelle prévue (RPP). Toutefois, à mesure que le nombre de strates augmente, cette solution devient rapidement coûteuse en raison de l'importance des calculs. La méthode proposée par les auteurs permet de réduire considérablement le nombre de calculs, moyennant la faible concession consistant à remplacer la RPP stricte par une RPP approximative.

Renssen et Martinus étudient l'utilisation de matrices inverses généralisées en théorie de l'échantillonnage. Après avoir passé en revue les propriétés des inverses généralisées, ils considèrent l'estimateur par régression généralisée lorsque la matrice des variables indépendantes n'est pas de plein rang et ils énoncent une condition de régularité aux termes de laquelle l'estimateur ne dépend pas du choix de l'inverse généralisée. Puis, ils présentent un algorithme pour le calcul des poids de régression et discutent brièvement de la pondération dans le cas de l'enquête sur la population active des Pays-Bas.

M.P. Singh

Dans ce numéro

Le présent numéro de *Techniques* comprend des articles traitant de sujets variés, y compris des exposés généraux sur la production de statistiques régionales, la qualité des données des bureaux de statistique, la non-réponse aux enquêtes et l'imputation, le plan de sondage, la collecte des données et l'estimation.

Dans le premier article, Brackstone énumère les stratégies adoptées par les bureaux nationaux de statistique pour mettre en place des programmes de statistiques régionales. La production d'estimations régionales sera le thème de plusieurs articles d'une section spéciale du numéro de juin 2003 de *Techniques d'enquête*. Pour commencer, l'auteur souligne le rôle essentiel des recensements et examine les questions associées à leur utilisation pour la production de statistiques régionales. Il mentionne d'autres sources éventuelles de données régionales, à savoir les fichiers administratifs et les enquêtes par sondage, utilisées isolément ou en combinaison aux données de recensement pour produire des estimations pour les périodes intercensitaires ou pour des variables que ne couvre pas directement le recensement. Il parle aussi des recensements continus, ainsi que des défis particuliers que pose la production de données régionales sur les entreprises et sur l'environnement. Enfin, il considère les problèmes d'organisation que doivent résoudre les bureaux nationaux de la statistique pour produire et diffuser des statistiques régionales.

Trewin passe en revue les pratiques et les stratégies adoptées par un organisme statistique national pour garantir la haute qualité des produits. Les bonnes relations avec les répondants, des employés compétents et motivés, des méthodes statistiques et opérationnelles solides et des programmes statistiques pertinents sont tous des ingrédients importants. Les défis qu'il faut relever à l'heure actuelle consistent à recourir davantage à des sources de données administratives, veiller à maintenir le bassin de connaissances et de compétences à mesure que les membres du personnel prennent leur retraite et répondent aux attentes grandissantes des utilisateurs des données. L'article est basé sur le discours principal fait par l'auteur au Symposium 2001 de Statistique Canada.

Thibaudau présente une méthode novatrice d'imputation de données sur les caractéristiques démographiques dans les enquêtes à grande échelle ou les recensements. Au lieu de s'en tenir à la méthode habituelle consistant à utiliser l'enregistrement complet le plus proche de compromis fondée sur traitement ou à créer des cellules d'imputation, il propose une méthode de compromis fondée sur l'estimation du maximum de vraisemblance d'après un modèle de probabilité conditionnelle. Cette méthode a pour but de créer des cellules proches, en ce qui a trait à l'ordre et aux caractéristiques géographiques, de l'enregistrement faisant l'objet de l'imputation. Thibaudau présente aussi une approche bayésienne intéressante d'évaluation de la méthode.

Nandram, Han et Choi considèrent le problème de l'analyse des données régionales multinationales en cas de non-réponse non-ignorable dans le cadre de l'inférence bayésienne. L'article représente une extension de travaux antérieurs de Stasy en supposant que les probabilités multinomiales suivent une loi a priori de Dirichlet et en utilisant une loi de distribution a priori sur les hyperparamètres. Les auteurs appliquent le modèle aux données sur l'indice de masse corporelle issues d'une enquête à plan de sondage complexe.

Dans son article, Stewart examine le biais que sont susceptibles d'introduire diverses stratégies de prise de contact utilisées pour réaliser les enquêtes téléphoniques sur l'utilisation du temps. Grâce à des études de simulation, il compare deux stratégies de prise de contact, l'une comportant un calendrier de prises de contact basé sur les journées convenables pour le répondant, où la journée de référence désigne un jour de la semaine, et l'autre, un calendrier basé sur des journées désignées, où la journée de référence reste fixe.

Bell et McCaffrey se penchent sur le problème de l'estimation non biaisée de la variance des coefficients de régression linéaire en cas d'échantillonnage à plusieurs degrés lorsqu'on ne sélectionne qu'un petit nombre d'unités primaires d'échantillonnage (UPF). Ils commencent par examiner les situations où le biais de l'estimateur par linéarisation de la variance peut être important, puis ils proposent un estimateur par linéarisation à biais réduit de la variance. En outre, ils utilisent une approximation de Satterthwaite pour déterminer le nombre de degrés de liberté qu'il convient d'utiliser pour les tests et les intervalles de confiance lorsque l'on se sert de l'estimateur à biais réduit proposé.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 28, numéro 2, décembre 2002

TABLe DES MATIÈRES

Dans ce numéro	123
GORDON J. BRACKSTONE	125
Stratégies et approches relatives aux statistiques régionales	125
DENNIS TREWIN	135
L'importance d'une culture de la qualité	135
YVES THIBAUDEAU	147
Imputation pour la non-réponse partielle basée sur un modèle explicite pour les catégories démographiques	147
BALGOBIN NANDRAM, GEUNSHIK HAN et JAI WON CHOI	157
Un modèle bayésien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions	157
JAY STEWART	171
Évaluation du biais lié à diverses stratégies de prise de contact dans les enquêtes téléphoniques sur l'emploi du temps	171
ROBERT M. BELL et DANIEL F. MCCAFFREY	185
Réduction du biais des erreurs-types pour la régression linéaire dans le cas d'échantillons à plusieurs degrés ...	185
MONROE G. SIRKEN	199
Effets de plan de sondage dus aux bases de sondage dans les enquêtes auprès des établissements	199
LOUIS-PAUL RIVEST	207
Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises	207
WILSON LU et RANDY R. SITTER	215
Méthode pratique de stratification multiple par programmation linéaire	215
ROBERT H. RENSSEN et GÉRARD H. MARTINUS	225
De l'utilisation des matrices inverses généralisées dans la théorie de l'échantionnement	225
Remerciements	231

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G.J. Brackstone
Membres	D.A. Binder G.J.C. Hole C. Patrick R. Platak (Ancien président)
COMITÉ DE DIRECTION	
Rédacteur en chef	M.P. Singh, <i>Statistique Canada</i>

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistique Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hildreth, *Statistique Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*
G. Nathan, *Hebrew University, Israel*

Rédacteurs adjoints

J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *Statistique Canada*
D. Norris, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.T. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Schuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliam, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Janvier 2003

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2003

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2002 • VOLUME 28 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



6153



NUMÉRO 2

•

VOLUME 28

•

DÉCEMBRE 2002

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



